No. 11

Data Processing Guidelines

Volume 1

BASIC DOCUMENTATION

MAY 1980

INTERNATIONAL STATISTICAL INSTITUTE
Permanent Office • Director: E. Lunenberg
428 Prinses Beatrixlaan
Voorburg, The Hague
Netherlands

WORLD FERTILITY SURVEY
Project Director:
Sir Maurice Kendall, Sc. D., F.B.A.
35-37 Grosvenor Gardens
London SW1W 0BS, U.K.

The World Fertility Survey is an international research programme whose purpose is to assess the current state of human fertility throughout the world. This is being done principally through promoting and supporting nationally representative, internationally comparable, and scientifically designed and conducted sample surveys of fertility behaviour in as many countries as possible.

The WFS is being undertaken, with the collaboration of the United Nations, by the International Statistical Institute in cooperation with the International Union for the Scientific Study of Population. Financial support is provided principally by the United Nations Fund for Population Activities and the United States Agency for International Development.

This publication is part of the WFS Publications Programme which includes the WFS Basic Documentation, Occasional Papers and auxiliary publications. For further information on the WFS, write to the Information Office, International Statistical Institute, 428 Prinses Beatrixlaan, The Hague – Voorburg, Netherlands.

Data Processing Guidelines

Volume 1

Prepared by:

WFS CENTRAL STAFF
International Statistical Institute
35-37 Grosvenor Gardens
London SW1W 0BS U.K.

Contents

Volume 1

Volume 2

Appendix II DP Manual For Processing Data from the WFS Core Questionnaire

Chapter 1 General System

1.1 Introduction

There are two major objectives in the data processing stages of a survey once office editing and coding are complete. The first is to "clean" the data by performing a series of comprehensive checks on its completeness and internal consistency, making appropriate corrections where necessary. The second objective is the production of analytical results, which involves the recoding of variables into the form required for analysis as well as the production of actual statistical tabulations.

This document discusses the procedures that are recommended for the computer processing of the data from the WFS household schedule and individual questionnaire. Included in it are recommendations applicable to the processing of any survey data as well as specific instructions for WFS data. Appendix I lists some general purpose software that is available for survey data processing. Appendix II gives complete specifications, documentation, input to and output from computer runs, etc., for processing a sample of data.

It is recognized that each country survey in the WFS may be different from any other in several aspects, e.g.:

- Different questions may be asked or the same questions asked but in a different order;
- Different coding, numbering, recoding, and tabulation schemes may be used; and
- Different software or hardware may be available for the processing of the survey.

As far as possible this document is survey-independent within the broad WFS framework and can be used regardless of survey design or computer availability in the country. Specific requirements are explained in detail.

In this document, reference is made to the following WFS documents:

WFS Core Questionnaire, March 1975
Modifications to WFS Core Questionnaire and related documents, June 1977
Editing and Coding Manual, May 1976
Guidelines for Country Report No. 1, December 1977
Users' Manual for CLUSTERS, June 1978
Users' Manual for WFS Date Edit, Imputation, and Recode Program, June 1980

1.2 Data and its Documentation

1.2.1 The Data

Data in WFS surveys are collected from a national sample of households. In each selected household, a household schedule is used to record basic information on general household

characteristics and on each member of the household. Much more detailed information is recorded for selected individual women in each household (or in a subsample of the households). In some surveys, the individual interviews are conducted at the same time as the household member listing. In other surveys, the individual interviews are made at a slightly later date.

For the individual interviews, a core questionnaire is used, modified from the *WFS Core Questionnaire* to allow for country specific circumstances. Optionally, a selection of questions from special modules (Family Planning, Abortion, Fertility Regulation, Mortality, Factors Other Than Contraception Affecting Fertility, Economic Data, Community Data) may be added to the questionnaire.

The data from the household schedule and from the individual questionnaires are transferred to a computer readable medium (cards, tape, disk or directly into the computer via on-line data entry) from coding boxes on the questionnaires, from coding sheets or a combination of the two. As all the data from one questionnaire will not fit on one card (where the word "card" is here used to mean the computer medium used be it actual cards or tape/disk records), the information must be punched onto several cards. Normally natural breaks in the questionnaire (e.g. sections) are used in designing the card layout to determine where a new card should begin.

1.2.2. Questionnaire and Card Identification

In order to recognize the different cards, each card is allocated a card type. Sometimes more than one card of a particular type may be required (e.g. WFS core questionnaire card type 4 for a woman's pregnancies). In this case a card sequence number is required in addition to the card type. Sometimes this card number is incorporated into the card type to give a two or three digit card type (e.g. WFS core questionnaire card types 41, 42, 43).

Each card must also carry a unique questionnaire identification. This will consist of a household number and the household member line number taken from the household schedule. The household number will either be a unique serial number (four or five digits long) or will consist of a cluster or sample area number and a household number within sample area (normally each three digits long making a six digit household identification).

1.2.3 Missing Data

Two special codes are normally set aside for all questions in any survey meaning, respectively, "respondent should have responded to the question but did not" and "this question was inapplicable and not asked". It is desirable that the same codes are used for each question throughout the data for these two meanings.

The conventions used in WFS data are that not stated (NS) responses are coded as fields of 9's. Not applicable fields (NA) are normally left blank at the coding and keypunching stages. However, the presence of any non-numeric character, blanks included, requires special treatment by computer programs. Therefore in final data files for analysis (cleaned and recoded data), inapplicable fields are coded as fields of 8's. This can be done while editing or recoding the data. There are, however, some advantages of doing it at the manual coding or keypunching stages, and these should be considered.

Not stated responses are sometimes imputed either manually or by computer to a "real" value. In the WFS data, this is done only in the case of missing date information in the marriage and birth histories of the individual women.

1.2.4. Data Documentation

(i) Codebook

The codebook is the document which describes the contents of each type of card, e.g.

Card type 51

Question	Description	Column	Codes
	Card type	1-2	51
	Cluster	3-5	001-135
	Household	6-8	001-150
	Line number	9-10	01-99
Q221	Did you breast-feed	11	1 yes
			2 no
Q222	Months breast-fed	12-13	0-76
			98 still breast-feeding
			99 not answered
Q223	No. of births	14	1 1 birth
			2 2 or more births
	"	"	"
	"	"	"

The codebook is similar to the coding manual given to the office coders during the coding process. However, it contains no coding instructions but does have a complete record of all questions, their locations on the cards, and all their possible values, including the meaning attached to the codes of categorical variables. This codebook must be prepared before starting to process the data by computer. Its preparation is a useful way for the data processing personnel to familiarize themselves with the data. (A complete codebook for the WFS Core Questionnaire is given in Appendix II).

(ii) Machine Readable Data Description

All general-purpose data analysis software requires a description of the data to be analysed. This data description consists, at minimum, of the location of the different variables in each type of record. More sophisticated packages provide for labelling of variables and variable categories. To do this, the information from the codebook must be supplied in machine readable form.

WFS has developed a standard format for a machine readable codebook called the WFS Dictionary. Once the data description is in this format, it can be converted by computer to the form required by a particular package, such as SPSS, COCENTS, CONCOR, etc. A program, CONVDICT, has been developed at the WFS headquarters to do this. The codebook for the "standard recode" data file in Appendix II is in WFS dictionary form.

1.3 Data Processing Steps

The data processing can be divided into five parts:

Keypunching — The data are transferred to a computer readable medium

Editing	The data are checked and corrected for • format and structure errors to ensure that all and only required data are present • out of range and inconsistent responses (imputation of missing date information may be necessary here)
Recoding	The edited data are transformed from the actual responses to a set of variables convenient for analysis
Tabulation	The recoded data are tabulated according to the specifications for country report No. 1.
Archiving and further analysis	The different data files with complete documentation are organized for further research.

1.3.1 Keypunching

Layouts of each card type should be given to the keypunching staff for easy reference to locations of the different fields on the cards. If normal card punches are being used, these should be set up with a program to copy the identification field on each card for a particular individual. This will avoid potential differences of identification between cards for one individual. All data entered must be verified.

If more sophisticated programmable data entry machines are being used, some data editing may be done at data entry stage. In particular, it is recommended that editing described under "format edit" in Section 2.2 be done.

1.3.2. Data Editing/Cleaning

It is important to the interpretation of the data that all possible errors and inconsistencies are corrected before the analysis phase. This cleaning or editing of data is an extremely important function involving both the demographers and data processors.

Data checking is usually done both manually in the office and by computer. Essentially the computer editing is a repetition of the manual editing and is necessary both because of human error in the manual operation and to correct errors introduced during coding and punching. The office editing procedure is described in the *Editing and Coding Manual* and will not be discussed further here. After the office editing, a more comprehensive checking must be carried out by computer. It should be heavily emphasized that data cleaning is not a trivial task. Once computer programs have been written or installed and tested the process of putting the data through these programs, using the output error lists to look up corrections from the questionnaires, updating the data and rerunning the checks is very time-consuming and will take from 3-12 months for an average size WFS national survey of 6,000 respondents.

Machine editing can be divided into two main stages:

(i) Format and Structure Checks

- Each card has a valid card type (including card number where applicable).
- Each part of the identification (e.g. sample area, household, and line number) contains a valid value.

- All fields contain numeric digits or blanks (alphabetic codes or other special codes, should never be used).
- Columns that should *always* be blank on a card (e.g. at the end of a card) are in fact blank. This detects cards where a shift in punching has occurred.
- All sample households are present.
- Cards for each household member are present (there are no gaps in the line numbers).
- All and only household members indicated as eligible have data for the "individual" interview.
- For individual data, all and only required cards are present for each respondent.

(ii) Range and Consistency Checks

- All codes are within the ranges specified for them in the codebook.
- All skips in the questionnaire have been correctly executed.
- Codes for filter questions that summarize previous information are consistent with that information.
- The information recorded is internally consistent.
- Dates in the marriage and birth histories flow in sequential order with a specified minimum elapsed time between events.
- Information on the household schedule is reasonably consistent with that on the individual questionnaire.

For some dates in the birth and marriage histories, only the year is given (or perhaps a "years ago" indication). Using an empirical or a statistical technique, missing months may be imputed. WFS recommends this procedure only for missing months of dates that cannot be otherwise obtained. All other missing information should be coded as "not stated".

The computer is used to locate errors and not to make corrections. During format, structure and consistency editing, error reports are produced from the computer. Correct values are looked up in the original questionnaires and written onto suitable update forms along with the identification of the record to be corrected. This work is usually done by the office editors/coders. It is therefore very important that:

- the error reports from the computer are clear and easily comprehensible to non-data processing staff

  and
- the update forms for writing down the corrections are simple to fill out.

Examples of an error report printout and of an update form are given in Appendix II.

Careful organization of the way corrections are done is also essential. Questionnaires should be easily accessible and located on shelves clearly labelled with the cluster/region to which they belong. The editing staff looking up the corrections must be thoroughly trained on how

to interpret error listings from the computer, how to look up appropriate corrections and how to fill out the update forms. The contents of update forms are keypunched and used to update the computer files. The whole checking and correction procedure must be repeated until no more errors are encountered.

The final report on the data should contain some quality-of-data measures. One of these measures can be derived from the number of errors found during the machine editing and a count should be kept of each type of error: field errors, office coding errors, keypunch errors, and errors caused by incorrect corrections. The staff doing the corrections should be given suitable sheets so that, for each correction, they identify the source of error and then mark the appropriate box, e.g.

Type of error	**Areas: 015-030**
Field	𝍸 III
Coding	𝍸 𝍸 IIII
Keypunching	IIII
Re-correction	II

1.3.3 Recoding and the Standard Recode File

Once the survey data file has been completely edited, a new file is created containing the actual variables that are to be used for analysis. The "recoded" data file created for producing Country Report No. 1 tables is known as the Standard Recode file. The layout of this file and the recoding operation are described in Chapter 5.

The information on this file does not correspond directly to the coded questionnaire: it is a summary of the information available in the questionnaire. The recoded file should, as far as possible, have the same structure in all countries as this simplifies the production of the tables for the Country Report No. 1 and later the cross national comparative analysis. However, since questionnaires vary from country to country, modifications to the recoding instructions as given in Appendix II will have to be carried out before starting to write the program for recoding.

In addition to the basic variables given in Appendix II each country will have asked important extra questions. These are added to the end of each respondent's record as "*country specific variables*". The file then captures nearly all the information from the questionnaire and can be used for most further analysis purposes. For ease of use Standard Recode data files should have associated machine readable WFS dictionaries describing their contents.

1.3.4 Preparation of Country Report No. 1 Tables

The minimum tabulations needed for preparation of Country Report No. 1 are described in *Guidelines for Country Report No. 1.* Appendix II specifies these tabulations in terms of the recoded variables in a form more convenient for the programmer. The layout of the different types of tables is explained in Chapter 6.

Country-specific tabulations may also be required and should be specified by the data analyst in a similar fashion.

Statistics on sampling errors are also given in the report. Details about their computation are given in Appendix II.

1.3.5. Archiving data and preparations for analysis

During the processing of a survey a number of computer tapes will be used, most of them as intermediary tapes. After the tables have been produced the following tapes should be kept for possible further processing and second stage analysis:

Raw data tape:	the original data as it was written to tape.
Structure tape:	after all structure and format errors have been detected and corrected.
Consistency tape:	after the consistency checking and correction have been completed (if imputation is done as a separate step, the tapes both before and after the imputation must be kept).
Recode tape:	after the recoding has been completed and checked.

All these tapes need complete documentation, details of which are given in Chapter 7.

Marginal (or frequency) distributions of responses to each question are a basic analytical tool useful at all stages of data processing and analysis. They should be produced:

- before the consistency checking, from the "structure tape" to give information about the quality of the data and to evaluate the need for imputation of missing information (e.g. incomplete dates);
- after the data have been edited, from the "consistency tape" to confirm that all values are now valid and to provide a reference document for the data; and
- after the data have been recoded, from the "recode tape", to confirm that the recoding was correctly carried out.

1.4 DP Strategy Considerations

1.4.1 Processing household and individual data together versus separately

Data from the household schedules is sometimes punched separately from the individual data and sometimes together, household by household. Either way, the two types of data can be sorted together into one file or separated into two files as desired. A decision needs to be taken at the start on whether to process the data separately or together. Examples of both methods are given in Appendix II.

The method chosen may depend on the software and hardware available. Some relevant considerations are:

- There is more likely to be software available for processing the separate files than

for the more complex combined file.

- If data are kept together and structured checked together each time the data is read errors of structure are less likely to be introduced when updates are made.
- Structure checking with two files implies matching them and identifying non-matches as one step in the process.
- Putting the data together may require sorting at an initial stage which may destroy the order in which the data were originally punched. This in turn makes errors in the punching of identification fields more difficult to locate.
- With separate files, less data has to be handled at once which may be an important consideration on small computers.
- Using one combined file is conceptually tidier and involves less record keeping and a fewer number of computer runs.

1.4.2 Rectangular versus structural files

A file is described as *rectangular* (or flat) when the same amount of data exists for each respondent (either equal number of cards per respondent or one long record). A structured file on the other hand may contain different numbers of cards for each respondent, e.g. in the WFS data, each woman may have a different number of birth and marriage history cards or a data file may contain all cards for complete households: household characteristics, household members, and eligible women cards.

In general, processing rectangular files is simpler. Indeed, much available general purpose software requires data in that form. Forcing data into rectangular form, however, may require padding out with a lot of blank data, e.g. if it is possible for a women to have up to 24 births, space must be left in every woman's data for that number.

Final data files for analysis should always be created in rectangular form, e.g. the WFS standard recode file. A choice may be made as regards the original raw (or card image) data. If thought desirable, rectangularization or padding in of cards containing inapplicable codes, can be done at the structure editing stage.

1.4.3 Availability of computer software

Some of the data processing can be done with existing and readily available software, although sometimes special computer programs will have to be written.

There are many packages available for data tabulations and slightly fewer for more extensive data analysis[1]. General purpose software for data cleaning purposes is not so plentiful. The available software varies greatly in its capacity, hardware requirements, ease of use, efficiency, flexibility, and standard of printout. Some suggestions for potentially useful software for both data cleaning and analysis are given in Appendix I.

[1] Rowe and Scheer, *Computer Software for Social Science Data,* Social Science Research Council, UK 1977.

1.5 Preparing for Data Processing

1.5.1 The DP Manual

It is strongly recommended that a complete set of documents relating to the data processing for the survey from the initial planning schedules and the basic documentation, such as the questionnaire, to listings of actual programs used, be kept in a single place (e.g. in a loose leaf binder). We will call this the DP manual. It serves two main functions:

- It is the working manual containing all necessary documents and specifications for the preparation of computer programs and for controlling the data processing.
- It will form, at the end of the DP phase, a complete record of all processing including the final specifications and listings of all programs (or control cards for package programs) used.

Appendix II is an example of a DP manual for a hypothetical survey based on the core questionnaire. Note that to avoid repetition with other WFS documentation, some sections contain only a sample of, or a reference to, what should be the contents. In addition, in a few sections, alternatives are given depending on the general strategy chosen. In a real survey, all sections should be complete and should reflect only what was actually used or actually done.

During the DP phase the DP manual is a working document and must be accessible to anyone involved with the survey. It is essential that it is updated as soon as any change occurs so that both during and by the end of the processing it truly reflects the processing steps used. After the first report from the data is published, reference to the manual will be necessary by researchers doing further analysis on the data in the event of queries on the meaning or consistency etc. of particular data elements. Thus the DP manual is an essential part of the archived data.

1.5.2 Planning

Initial data processing plans are drawn up when the survey is first designed and the project budget calculated. At this time, the various tasks to be performed are identified and a flowchart of the data processing steps is drawn. Existing software is evaluated in the context of these tasks and decisions are made whether to use procedures from a general purpose software package, whether to modify existing special purpose programs or whether completely new programs need to be written. Some suggestions on available software are given in Appendix 1.

Estimates of human time in person days and also of elapsed time are then made for preparing documentation, for writing and testing programs and for performing the actual data processing. The computer time required is also estimated. The list of tasks to be performed, the flowchart for the data processing, and the bar charts for the work comprise the first documents to be entered in the DP manual. Examples are shown in Appendix II.

1.5.3 Preparing Documentation and test data

As soon as the questionnaire is finalized, a code book must be prepared. The codebook describes the contents of each card type giving for each question, the columns it occupies, the possible codes or range of codes and their meanings, and details of any special codes (see section 1.2.4. above).

The questionnaire and the codebook form the basic documentation for the Survey data and are the next items to be placed in the DP manual.

Specifications for the various DP stages are then written and checked in collaboration with other survey team members. These include keypunching specifications, machine editing specifications, recoding and tabulation specifications. They will not all be prepared before data processing starts but must be available well before programming starts for the corresponding phase. As they get completed, they are entered in the DP manual.

A set of test data of about 50 cases is required. The simplest way to make this is to fill out 50 questionnaires and then code them. While doing this, as many of the different errors as possible should be deliberately introduced by referring to the machine editing specifications.

1.5.4 Program Specifications

For each program that is required according to the flowchart of the data processing, complete specifications must be available before starting to do the programming. These require:

- a description of the purpose of the program.
- a diagram showing the inputs and outputs.
- a completed description of the contents of the input and output files including details of any printed output.
- a description of all possible error messages.

1.5.5 Program Writing and Testing

Program writing will mean one of the following:

- Preparation of control cards for a procedure in a general purpose package.
- Modification of an existing program.
- Writing of new programs.

If new programs are going to be written for a majority of the tasks, then a particular strategy is recommended. This involves developing a basic program to read the survey data and to detect and print basic structure errors. Once this program has been developed, it can be used as a starting point for any other programs required, including complete structure checking, machine editing, and variable recoding.

This approach offers two main advantages. First, the data file will always be accessed in a consistent way including a basic structure check, so that any updates which by mistake destroy the basic structure will be detected. Secondly, the existence of a complete, tested program for reading the data file will allow the programmer to concentrate on the specific application, be it consistency checking or recoding, which will facilitate program development. The program specification in Appendix II for the program STRUCT4 describes this technique in more detail.

It should be stressed however that the use of existing general purpose programs is nearly always preferable, being much less time consuming and error prone than starting programming from scratch, even if the way the processing is done has to be slightly tailored to the software (e.g. using two different programs where one specially written one could have done the task).

The previously prepared test data is used to test all programs. This includes testing the use of existing general purpose procedures. After individual testing of programs, the test data should be passed through the entire data processing scheme. This will include doing corrections to the data where necessary before proceeding to the next step. The output from these test runs should be filed in the DP manual as proof that the programs work and as an easy reference later on how the program is used and an example of its printout.

1.5.6 Data Processing Organization and Control

It is recommended that keypunching, machine editing, and correction be done in batches, for example, region by region, as the data becomes available. There are various reasons for this:

- systematic errors in office editing, coding or keypunching can be detected early, while processing the first batch of data.
- editing and correction are done on a manageable number of questionnaires at any one time.
- time is not wasted waiting for all data to be ready before starting complete processing.
- checking listings of data against listings of households interviewed in the field is easier on a small set of questionnaires.

In some WFS surveys, data from the household schedules is punched and processed separately from the individual data and in other surveys all data is processed together. The strategy chosen will affect the organization of the work. Both methods are discussed in the rest of this document. However, in general if the household data is processed separately, it is only the individual data that is processed in batches. The household data would be edited as one complete file.

During the data processing, it is important to know exactly which stage each batch of data has reached. To this end, a data processing control sheet should be designed which lists each task (using the flowchart from the planning stage) and leaving room for the date on which it is performed. The tasks may be grouped into the various phases (e.g. structure editing). Since

each phase will go through several iterations before it is complete, several dates for each task will be entered, one for each time it is performed. When the phase is complete, the completion date is noted and the batch of data moves into the next phase. Such a control sheet will be needed for each batch. Once the batches have been merged, a single control sheet for the later data processing stages will be used. Sample control sheets can be found in Appendix II.

Processing data in batches requires careful control of tape (or disk file) usage. It is suggested that the tape number (or disk file name) and number of records are noted on the DP control sheet each time a new file is created.

Chapter 2 Format and Structure Edit

2.1 Overview

Invalid card formats and incorrect file structure make all subsequent data handling very difficult. Errors in file structure should therefore be located and corrected as the first stage in the editing process.

The general data processing system for this is shown in Figure 2.1 Here is a summary:

The cards are put to tape (or disk) and optionally listed.

The data are run through the format check program. Errors are corrected and the format check repeated again until all errors are removed.

The unsorted data tape is sorted into ascending order using the questionnaire and the card identification as sort keys.

Structure editing is performed on the sorted data tape until all errors found have been corrected. At the end of this step, two files exist: one containing all household and household member cards and the other containing all cards for individual interviews.

Marginals for each question are produced from the structurally correct files to give an initial idea of the quality of the data. These marginals will quickly show which variables have out of range codes by comparing them with the codebook.

A tabulation or print program is run to check that the correct sample households from each cluster or area are present.

2.2 Format Edit

This is the first data processing operation carried out on the survey data and has four components: card type, identification, blank columns and numeric checks.

2.2.1 Card Type Check

The card type in each card is matched against a list of valid types.

If a card type can have sub-card types (e.g. card numbers), these are matched against a list of valid card numbers.

2.2.2 Identification Check

The range of each part of the identification field is checked for validity. In some surveys it may also be possible to go further and check, say, the range of valid household numbers within each sample area.

For example:

Sample area	Range of household numbers
001	01-49
002	01-73
003	02-61
etc.	etc.

2.2.3 Blank Columns Check

This is to see if shifting has occurred during keypunching. The check is that all columns that should be blank on a particular card type are blank. For example, in the core questionnaire columns 59-80 of card type 3 should always be blank.

2.2.4 Numeric Check

All non-blank columns should contain a number (i.e. digits 0-9).

2.2.5 Error Reporting

The errors found should be reported on an error list, which should (at least) contain the following information:

- Sequence number of the card within the file. (This is so that updates can be made by overall sequence number since the file has not yet been sorted).
- The card in error.
- An indication of the error, which can be a message printed after the card, asterisks (***) printed under the error, or any method normally used in the DP installation concerned.

The format edits should be run until no more errors are found.

2.2.6 Conversion of Blanks to Inapplicable Codes

At this stage blank columns (excluding those columns at the end of cards which are never punched) may be converted to inapplicable codes for the corresponding questions (normally 8's). The reasons for doing this were outlined in Section 1.2.3.

2.3 Sorting and Separating the Data

Before checking the structure, the data may be sorted so that all cards belonging to one household are together and also so that cards with duplicate questionnaire identification and card type fall together and can be detected. It should be remembered that these duplicates may indicate mispunching of the identification or card type rather than genuine twice-punched data. They may therefore be taken as a warning of gaps in other questionnaires.

The cards should be sorted into ascending order by household identification (which may consist of several fields, e.g. area, household), line number and card type. Most computer installations have a sort program that can do this.

If the data was originally punched in the required sort order, it may be preferable to sort *after* the structure checking step. In this way, cards with mispunched identifications are more easily identified.

If the structure checking is to be carried out on the household and individual data separately, then at this stage the data should be separated into two files:

- household characteristics and household member cards from the household schedule — *the household data*
- cards from the individual interview — *the individual data*

2.4 Structure and Completeness Edit

The structure and completeness edit consists of five checks

(i) All household member cards that should be present for a household are present if the "outcome" of the household interview was successful.

(ii) All household members indicated as eligible for the individual interview meet the criteria of eligibility.

(iii) All sample households have at least a household card. The total number of households, household members and selected women for each area agrees with the field work totals.

(iv) For each individual interview with a successful outcome all relevant cards are present.

(v) All women in the household eligible and selected for interview have individual data and only eligible women have this data.

If the data have been split into a household file and individual file, then checks (i)-(iii) will be done on the household file, check (iv) on the individual file and check (v) is made by matching the two files after they have been corrected for all other errors.

2.4.1. All Household Members for Each Household Present

- The line number for the household member cards should be in ascending sequence with no gaps and no duplicates and starting with line number 01.
- One household card must be present for each set of household member cards.
- The number of household member cards present for a household should equal the number of members in the household given on the household card.

2.4.2 Eligibility for Individual Interview

A household member is defined as eligible for the individual interview according to certain criteria for the survey. For example, to be eligible these conditions might be required:

residence:	lives in this household
sex:	female
age:	15-49
marriage status:	has been or is in a union

On each household member card, there is a field indicating whether the person is eligible or not. In addition, since in some surveys not all eligible women are selected for interview, there can also be a field indicating whether an eligible household member was selected.

The following consistency checks should be made on the household data:

- Household members indicated as eligible are eligible according to the survey criteria.
- Household members indicated as ineligible are not eligible according to survey criteria.
- The number of members indicated as eligible equals the check value given on the household card.
- The number of members indicated as selected is equal to the check value given on the household card. (This check is only done if not all eligible women are selected).

2.4.3 All Households Present

The following totals should be produced for each cluster or area and for the total sample and checked against the records for field work, office editing and coding:

- Number of households.
- Number of households successfully interviewed (using the result code from the household card).
- Number of eligible women.
- Number of eligible women selected for interview (where applicable).
- Number of selected women successfully interviewed.

If a simple tabulation program is available, the first two are given by a table of cluster by result code from the household cards and the rest from tables of eligibility, whether selected, result code by cluster from the household member cards, e.g. for number of eligible women:

Cluster	**Eligible**	**Not Eligible**	**Total**
1			
2			
3 : : :			
Total			

As a more complete check, a listing of all the household and household member cards can be made and the identifications checked against the field works sheets. This will also catch errors in the identification field. It is recommended that such a listing be made for each cluster or area separately and with 2-3 blank lines between each household so that it is easier to read.

2.4.4 All Relevant Individual Cards Present

Each individual interview is punched onto various different card types. It must be checked for each interview that:

- Each mandatory card type is present
- There are no duplicate cards
- One and only one of a set of alternative versions of a card type are present
- There are no invalid card types (these should already have been corrected for during the format edit)
- Where multiple cards of the same type exist (e.g. pregnancy history), that they follow sequentially and that there are the correct number according to a check field on some other card. (This latter check alternatively may be left until the consistency check phase of editing).
- Where the presence of a card is dependent on another card also being present, that this condition is satisfied.

In preparation for this check, the rules for the presence of cards should first be specified, e.g.,

Card type	Requirement	
20	Mandatory	
30	Mandatory	
41	Optional	
42	Optional — if present, 41 must be present	
50	Mandatory	
60	Mandatory	
70	Mandatory	
81	Optional	One and only one of these must be present.
82	Optional	
83	Optional	

The program should print out all cases which do not follow these predefined rules.

If a full set of cards is not present for a case, it may be a genuinely "incomplete interview" as given by the "final result" variable on the first card for the case. All cards for individuals with result code other than "completed interview" should be deleted from the individual file, but kept in a separate "incomplete interview" file.

At this stage it is possible to rectangularize the file by ensuring that each individual has the same number of cards. This is done by padding dummy cards containing the correct identification and card type but with inapplicable codes (8's or blanks) for all other fields. The advantages and disadvantages of working with rectangular files for both consistency editing and analysis were discussed in Section 1.4.2.

2.4.5 All Eligible Woman Interviews Present

The final checks on the structure should be that:

- All household members indicated as being eligible (or selected where applicable) do have a set of individual cards.
- Incomplete interviews for eligible women do not appear on the individual file (these should have been eliminated during the "relevant individual interview cards present" editing).
- There are no individual interviews for non-existent household members, i.e. a set of individual cards present but no corresponding household member card.

If the structure checking up to this point has been done on the household and individual files separately, this last check can be done by matching eligible household member cards from the household file against the first of the set of cards for each woman on the individual file. All non-matching cases on either file should be listed and the appropriate corrections made.

2.5 Preparations for Consistency Checking

It is important now, before starting the complete consistency edit, to have a preliminary look at the contents and quality of the data. This can be done by producing simple marginal frequency distributions on all variables from each card type.

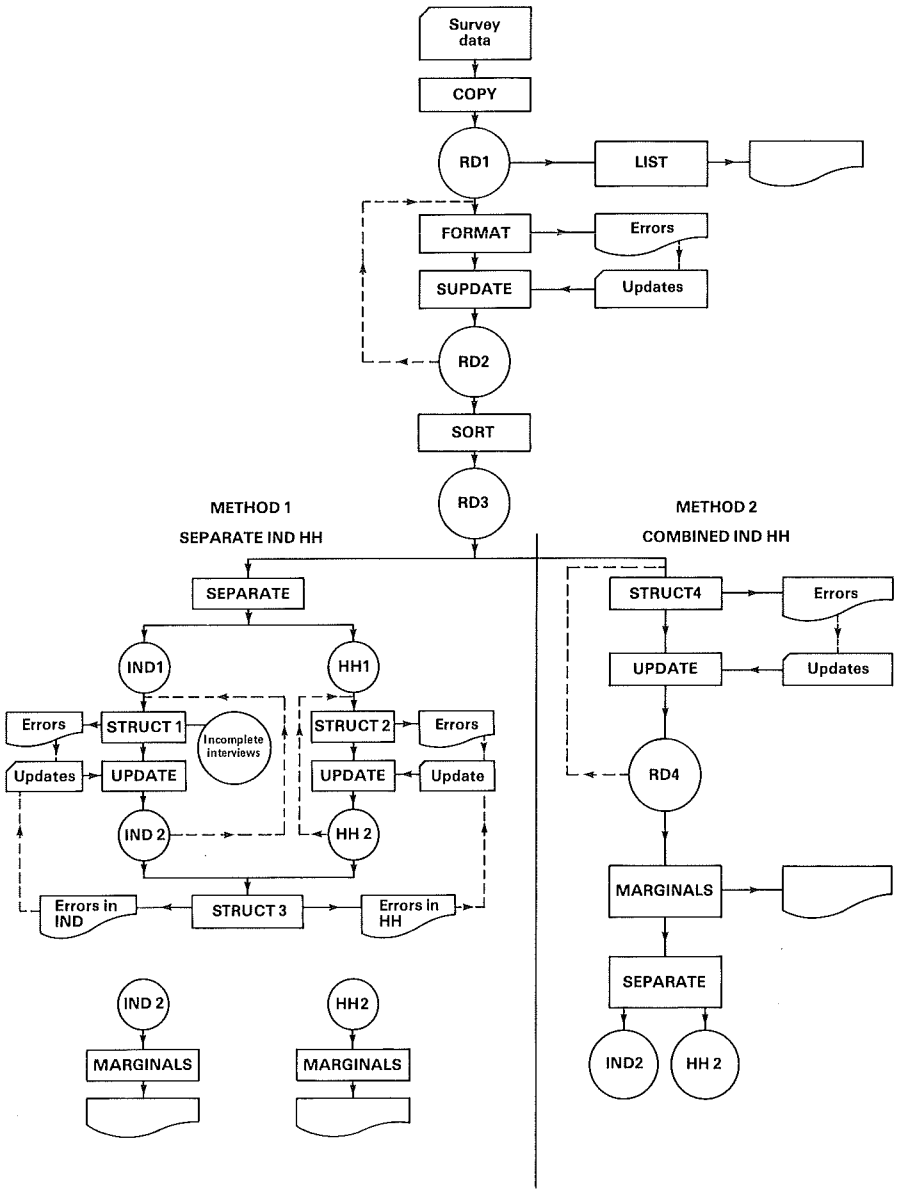
2.6 Summary

A flowchart of the format and structure editing stages is given in Figure 2.1. Eleven different programs are used in this figure.

Note that either STRUCT1, STRUCT2, STRUCT3 are required *or* STRUCT4, depending on whether structure checking is done on separate household and individual files or on the combined file.

FORMAT	— check format
SUPDATE	— update file by record sequence
UPDATE	— update file by record identification
SORT	— sort (normally available at the computer installation)
SEPARATE	— select cards and write to a separate file
STRUCT1	— check individual data structure (see 2.4.4)
STRUCT2	— check household data structure (see 2.4.1, 2.4.2, 2.4.3)
STRUCT3	— match household and individual data (see 2.4.5)
STRUCT4	— structure checks on combined household and individual file
LIST	— list selected cards
MARGTABS	— produce marginal distributions and cross-tabulations

Figure 2.1. Format and Structure Edit



Chapter 3 General Consistency Edit

3.1 General Discussion

3.1.1 Introduction

After the office editing has been completed (see *Editing and Coding Manual*) some of the information on the questionnaires will still be incorrect; also it is very likely that errors were introduced during the coding and punching of the data.

The discussion in this chapter assumes that:

- The files have the correct structure.
- The survey data file has been split into two files containing the household data and the individual interview data, respectively.
- All partially completed interviews have been dropped from both files.

3.1.2 Types of Checks

Before starting to edit the data, the editing rules must be carefully specified for each of the following types of checks:

- Range checks — that all questions have valid codes (Household and individual).
- Skip checks — that the skip pattern of the questionnaire has been correctly followed (Household and individual).
- Filter checks — that the questions that summarize previous information i.e. the "filters" are consistent with that information (Individual only).
- Table checks — that lines in a birth or marriage history table are coded without gaps, and that the expected number of lines are coded (Individual only).
- Miscellaneous checks — that the values in various fields are consistent with one another (Household and individual).
- Date checks — that dates of events are consistent (Individual only).

In order to deal successfully with the range, skip and filter checks, a network diagram (or flowchart) should be prepared for the individual questionnaire. This diagram is presented in Section 3.2. A flowchart of the entire individual core questionnaire is available in Appendix II. The checks for the household data are simpler and do not need this kind of flow chart.

3.1.3 Programming Strategy

It is recommended that the editing for the individual data be carried out in the following four different stages:

- The range checks;
- The skip, filter, and table checks;
- The miscellaneous consistency checks;
- The date checks.

Four separate computer programs may be used for these four stages. Alternatively, if hardware resources are large enough, all checks except the date checks, can be combined into one program provided that the logical sequence of the checks is maintained.

The advantage of the separate program approach is that the programs will be simpler to write and faster to run. The disadvantage is that each run may produce error reports of different kinds for the same questionnaire. The office editors thus may have to go back to the same questionnaires for corrections several times. On the other hand, each type of error report will have only a limited number of types of errors and will be much easier to follow. The data processing personnel must therefore decide in collaboration with the editors whether to process the different checks separately or whether to try and integrate them.
All checks for the household data can be made in one program.

3.2 Diagrammatic Representation of the Questionnaire

A diagram of the questionnaire is drawn to show the valid codes for each question and the conditions under which one or more questions may be skipped.

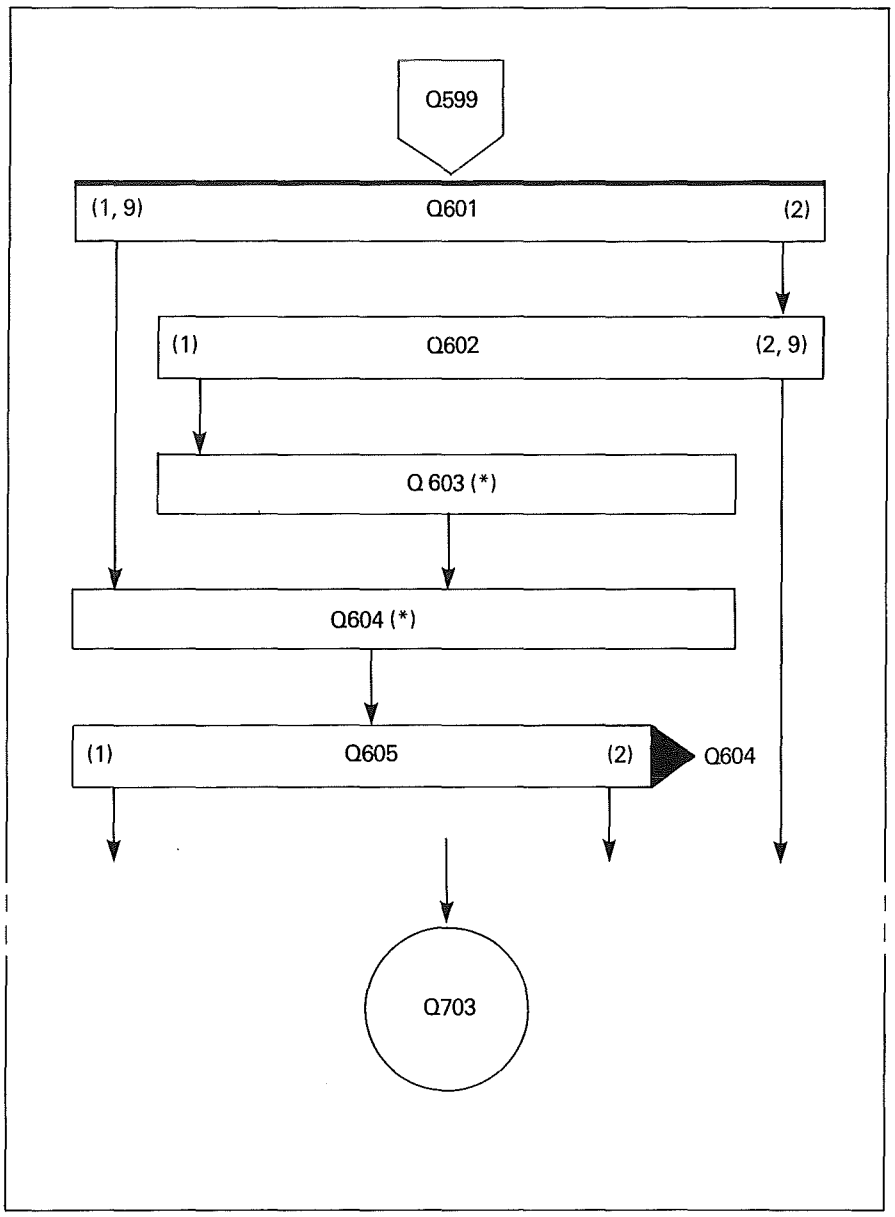
In the case of complex questionnaires (like the WFS Core Questionnaires) the diagram should be divided into sections. For each section a network diagram is constructed that shows the range of answers to each question, together with the skip and filter patterns. These aspects of the network diagram are discussed in detail below.

Once the network diagram for the entire questionnaire has been prepared and carefully checked the required range, skip, and filter checks are readily obtained. The pictorial representation becomes particularly relevant for sections of the questionnaire which have a complicated skip pattern. The structure becomes visually clear as the diagram is completed.

3.2.1 The structure of a network diagram

The diagrams consist of a series of boxes enclosing question numbers and range values, with branches connecting the boxes. Figure 3.1 highlights all features of the diagram taken from Section 6 of the Core Questionnaire.

Notes Figure 3.1. Network Diagram



Notes

(i) This notation is used to indicate which question(s) immediately precede the first question on this page; in the example, Q599 precedes Q601 unconditionally (it is possible to specify more than one question; each question can have a condition attached — see Appendix II for further examples.)

(ii) All questions with a heavy line on top of the box are always asked of all respondents; this means that the NA (Not Applicable) code is not valid for these questions.

(iii) In this example (which also applies to Q601 and Q605) the answer to the question determines the next question to be asked. The complete range of codes (except the NA code) is contained in the "question box".

(iv) This question (and it also applies to Q604) is an example where the answer to the question has no bearing on the next question to be asked, i.e. regardless of the answer to Q603, Q604 will always be asked (provided, of course, that Q603 was asked in the first place).

(v) The asterisk (*) indicates that the range values for this question are determined in the country.

(vi) The horizontal "arrow" at the right hand of the box is used to indicate that the filter question Q605 derives its answer from Q604.

(vii) This notation is used to indicate that the next question in sequence is to be found on another page; this notation is usually used between sections, which are normally coded on separate pages.

3.2.2 Ranges

The diagram does not contain the NA code for any question. In boxes with a heavy line on top this code is not allowed and should not be accepted. For all other questions the range checks must include the NA value as a valid code.

The range of numeric codes is specified in two ways (it is assumed that only numeric codes are used):

- If there is only one branch from a box, the values are entered immediately after the question number.
- If there are two or more branches from a box, the values are entered in the box above the branch that is taken when the question has that value (see the example above).

The NS (Not Stated) value is always entered where applicable. It is coded as one or more 9's, depending on the width of the field, i.e., 9, 99, 999, etc. The NS value is not a valid code for filter questions. There are also a number of other questions where the NS value should be considered invalid e.g. in the WFS Core Questionnaire Q108, Q201-Q213.

3.2.3 Skips

For every question where NS is a valid code, appropriate skip patterns for this response must be established and included in the diagram.

3.2.4 Filters

Filter questions have to be verified against the questions from which they are derived. They are distinguished by an arrowhead on the right hand side of the box containing the question. The question(s) from which the filter is derived are written immediately to the right of the arrow.

3.3 Range Checks

If marginal distributions for all questions have been obtained as recommended in Chapter 2, then these can be examined to find which questions have out-of-range responses according to the specified allowed ranges. The range checking program need then only check for and locate cases with out-of-range values on these specific questions. This should make the program faster to write and faster to execute. If however the range check program is prepared before marginals are produced, then it must check *all* questions.

When specifying the range checks, the points below should be taken into consideration:

- Except for questions that are applicable for all respondents the NA code should be included in the valid range. This may be easier if not applicable responses have been coded as a numeric value or if blanks have been recoded to a numeric value during the format checking as suggested in Chapter 2.
- For a large number of questions the NS code will need to be included in the range. Nevertheless it will be useful to omit it for the first set of range specifications. This approach would flag all missing data as errors; each offending case can then be examined to see whether the missing information can be obtained from other data on the questionnaire. Cases with missing *date* information in the birth and marriage histories should not be listed but treated later during date editing.
- There are certain questions where the codes can theoretically be in a rather wide range, but where a vast majority of cases fall in a narrower range. In such cases it may be more effective to specify initially a narrower range as the legitimate one, even though in this case some errors indicated will turn out to be not genuine on variable examination. An example of this implausibility checking is for the variable "age at first birth". This may generally be in the range (15-45), but a few legitimate values may be as small as 10 or as high as 49. Rather than take the range (10-49), take it as (15-45) to start with and make sure the legitimate but unlikely values are not errors of coding or punching.

3.4 Skip Checks

Skip checks are used to check that the correct set of questions are asked of each respondent. They follow directly from the network diagram. Questions which are applicable to all respondents need no skip checks since range checks should have verified that they are never coded as NA. Each potentially skipped question will require a two directional check i.e. that the question is not coded NA (not applicable) *if* and *only* if previous variables are coded in a specific way.

The code checks of this type require little programming expertise, but it must be clearly understood that the skip checks involve a *two-way* relationship. As an illustration of the skip checks for the part of Section 6 of the Core Questionnaire given in Section 3.2.1., we have:

Variable	Variable ≠ NA if and only if
Q602	Q601 = 2
Q603	Q602 = 1
Q604	Q601 = 1 or 9 or Q602 = 1
Q605	Q604 ≠ NA

3.5 Filter Checks

Filters are coded by the interviewer to guide the flow of questioning on the basis of information previously obtained; e.g. a filter question on whether a woman has used any family planning method, will be given the answer "yes" if at least one specific method has been marked as used. NS should never be a valid code for a filter question.

A safe way of developing a complete set of filter checks is to take all possible values for a filter and, by looking at previous questions, find out how they were arrived at. The filter checks are all of the form

If filter = a THEN b
where "a" is a numeric code in a filter, and
"b" is the required condition in terms of the source question(s).
e.g. for the family planning example given above:
If Q503 = 1 THEN Q315 = 1 or Q316 = 1

3.6 Birth and Marriage Table Checks

When dealing with tables, e.g. the birth history table, it should be checked that:

- The lines are coded without gaps, and once a line has been used there is no missing information.
- The number of filled-in lines is the expected number as given by a check question.

In the WFS Core Questionnaire, these checks apply to birth tables, "other pregnancy" tables and marriage tables. There are questions in the questionnaire on the total number of children and the number of boys and girls both dead and alive which should be checked against the number of entries in the tables.

3.7 Miscellaneous Consistency Checks

Every questionnaire will have many other consistency checks that can be performed across questions, e.g. number of family planning methods known is equal to the number specifically mentioned, or if a woman is sterilized then she should know of sterilization. A suggested list for the WFS Core Questionnaire is given in the data processing specification in Appendix II.

3.8 Date Editing

It is assumed that basic range checks on all dates are done during the range checking. However, since consistency checking of dates, especially those in the birth and marriage histories, is somewhat complicated, it is recommended that this be done as a separate procedure. This is discussed further in Chapter 4.

3.9 Conculsion

Each time corrections are made, the checks must be rerun to make sure that no new errors have been introduced. In addition, it is good practice at the end of each step to repeat the processing of the previous step, e.g. when the skip errors have been corrected, the range checks should be rerun.

When all the steps specified in this chapter have been performed (both on the individual and the household file) the survey data file should be clean except for possible errors in the temporal variables.

Chapter 4 Editing of Birth and Marriage Histories

4.1 Required Basic Checks

After the consistency edits and associated corrections described in the previous chapter have been carried out, the next stage of the data processing consists of the editing of birth and marriage histories.

Editing of *birth histories* essentially consists of checking that the births are in the correct historical sequence, that there is an acceptable minimum interval between successive births, that dates are consistent with the age of the respondent, and that births with no information on the year they occurred (in whatever form) are identified.

For *marriage histories*, editing consists of checking that marriages are in correct historical sequence (i.e. they do not overlap except in polygamous countries) and that dates are consistent with the age of the respondent.

More specifically, edit checks for the birth history are:

- The date of interview is not before the date of birth or death of any child.
- The date of death of a child is greater than the date of birth.
- The date of beginning of the current pregnancy is not before the date of the last birth.
- The age of the respondent at the time of giving birth to her first child is not less than some specified minimum.
- The time interval between neighbouring births is not less than the biologically possible minimum (say 8 months). This check assumes that births in the history have been recorded in chronological order.
- Further constraints as seem necessary, e.g. if the date of the woman's sterilization is available from the Fertility Regulation Module (Q571), then this date must not be before the date of the last birth.

For the marriage history, the checks are:

- The date of interview is after the date of the beginning and before the end of the last marriage.
- The age at first marriage is not less than the culturally determined minimum possible age at marriage.
- Events in the marriage history (or in polygamous countries, dates of marriages) are in chronological order.

4.2 Editing of Histories with Incomplete Data

A comprehensive edit procedure must take into account the problem of missing data. The procedure should be based on realistic assumptions regarding the form and completeness of data encountered; thus the following data defects should be provided for:

- For a significant proportion of the events, only the year but not the calendar month is available.
- For most of the events, information on the year is available, but this information comes in a number of alternative forms, e.g. as calendar year, as duration before the interview, as respondent's age at the time of occurrence, as duration from some other event, etc. and must be transformed into a common form before comparison.
- For a certain proportion of events no information about the year in any form is available. Such cases have to be identified and in general, a year has to be *imputed manually* before further processing of the data.

Hence the procedure will assume that the information on the month may or may not be available, but that the information on the year is either available in some form or has already been identified as missing and manually imputed.

4.3 Different Forms of the Dates and The Century Month Code

Dates of the various events may appear in different forms as:

- Calendar dates, i.e. actual month and year of an event.
- Duration before the interview, e.g. years ago or current age of child instead of birth date.
- Duration before or after another event, e.g. the duration of a marriage.
- Age of respondent when the event occurred.

In all cases care should be taken to establish from the field work team whether "years" are given as rounded or completed years.

In comparing dates, it is necessary to transform them all into the same form. The form used in the WFS is the century month code. The century month code is defined as the number of months after December 1899 to the date of occurrence of the event. If Y is the last two digits of the calendar year and M is the calender month, then:

$$CMC = 12*Y + M$$

For example, for March 1930:

$$CMC = 12*30 + 3 = 363$$

Note: to compute the year Y given CMC, the correct formula is

$$Y = (CMC-1) \div 12$$

4.4 Date Imputation

After all obvious inconsistencies have been removed by reference to the original questionnaires, we still have the problem of missing months.

It may seem undesirable to impute dates when months are missing for a large proportion of the events since it involves invention of data. On the other hand, if all but a small proportion of birth and marriage histories are complete (that is if century-month codes of nearly all events are recorded and consistent) then it may be tempting to avoid the missing value problem by rejecting incomplete histories. However, except perhaps for the most extreme cases on either side, it is recommended that the problem of missing months be handled by following a logical and consistent imputation procedure. The recommendation is based on the following considerations:

(i) Even when months are missing for a relatively small proportion of the events, it will be in general an unacceptable practice to reject incomplete histories, since the complete interviews alone may constitute a biased sample. For example, if dates of distant events are less well remembered, then respondents with longer birth or marriage histories will be under-represented in the reduced sample.

(ii) In the other extreme case, where months are missing for a large proportion of the events, the problem must be resolved somehow since outright rejection of offending cases is obviously no solution. It should be noted that month imputation has little effect on "broad" temporal variables such as respondents' age in five or ten-year groups, number of children born in the past five years etc., which could be calculated even if little information on months was available. In contrast, "fine" temporal variables, such as interval between marriage and first birth, are more sensitive to month imputation. However, these variables are normally used much less frequently in tabulations than the "broad" variables.

(iii) An apparently simpler procedure in which, for example, age is calculated by subtracting the calendar year of birth from the year of interview, nevertheless implies month imputation of a kind. Further, the fact that dates or events may be recorded in various forms (as calendar dates or as duration,) — and more variably in those surveys where months are more difficult to obtain — will tend to make alternative procedures for implicit or explicit month imputation quite elaborate.

(iv) At the recoding stage it may be decided that certain variables which are sensitive to month data need not be calculated because their quality is too low. The imputation procedure can give information to help in making the decision. Generally, the prevalence of missing months and the quality of the data with months present determine the *use* which is made of the data after the imputation stage, but they do not affect the imputation procedure itself.

(v) An alternative to imputation of months is to leave an NS code in all month variables where the information is missing. This may lead to complications in later analysis of the data, but the problems so created are not insuperable.

There are various imputation methods that can be used. For example, imputation could be done by assigning a random month to each date where only a year is given. A better method is to construct possible minimum and maximum dates, known as the logical range for each event, and to choose a point randomly in this range. The WFS imputation program mentioned in Section 4.6 uses this method.

Note that imputation often starts at a much earlier stage than the machine editing of the data. Interviewers and office editors may well do some guessing of dates before the data gets to the computer.

In conclusion it should be emphasized that imputed values are never a good substitute for data obtained in the field, and fabrication of data should be avoided whenever possible. Also, if data are machine imputed, then the original data before imputation should always be preserved.

4.5 Results from Date Editing and Imputation

After completing the date editing and imputation a complete set of dates (in the form of century month codes) for each event should be available. This constitutes most of the information needed for the first part of the "standard recode" data file, the construction of which is the next stage of the process.

4.6 The WFS Date Edit, Imputation, and Recoding Program (DEIR)

A program has been developed by the WFS to do a comprehensive check on all dates, to impute missing dates if required and to construct all variables of the first part of the Standard Recode data records (V001-V306). Complete details of the methods used for constructing logical ranges of dates, of applying constraints to them, of identifying inconsistencies and the ways imputation can be performed can be found in the user's manual for this program*. Note that use of this program requires that a special file of date data be first extracted from the individual data. Specifications for this "extract" are given in Appendix II.

4.7 Summary of Date Edit, Imputation and Recoding

A flowchart of the date edit, imputation and recoding process is given in Figure 4.1.

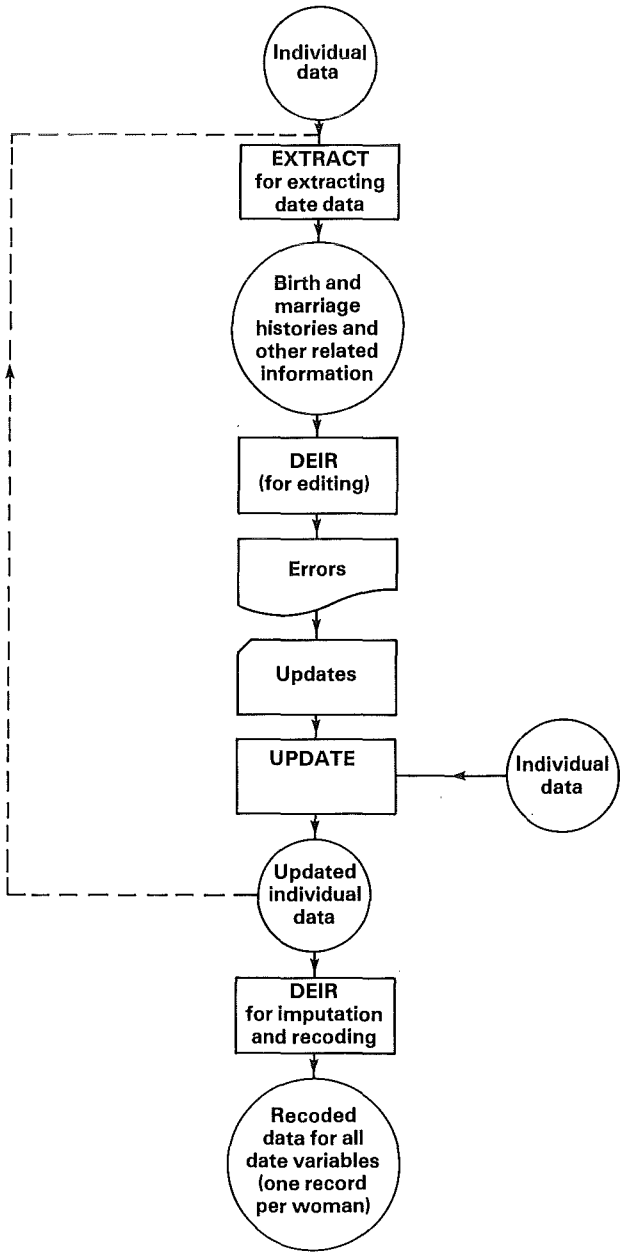
A file of date information is "extracted" from the individual data and passed through a special date editing and consistency check program. Corrections for reported errors and inconsistencies are found in the questionnaires and applied to the individual data file. Date information is extracted again and the checking and correction process repeated until no further errors are found.

Missing months are then imputed and a new file is created consisting of one record per woman, containing all the date-related information. To save time at the recoding stage, this output file should preferably be in exactly the same format as the first part of the WFS standard recode file described in appendix II.

The program "DEIR" has been developed at WFS headquarters to do these three tasks of date checking, imputation, and production of an output file.

*DEIR User Manual, WFS Tech. No. 1430.

Fig. 4.1 Flowchart for the date edit, imputation and recoding procedure



Chapter 5 Recoding of Variables

5.1 Reasons for recoding

The individual questions asked in a survey often do not correspond one-for-one to the variables that are required for analysis. For example:

- Educational level may be asked through 3 questions:

  Have you ever been to school?
  If yes: what level did you reach?
  what grade in that level?

  For some analysis purposes, a *single* variable giving years of education based on the answers to these 3 questions might be required.
- A question may be asked which gets a numerical response, e.g. age in years, which for tabulation purposes one might require to collapse into a few groups.
- Individual questions on a set of items might be asked, e.g. whether each of a series of specific family planning methods have been heard of. For later tabulation purposes one might want to reduce these to a single variable indicating whether *any* method has been heard of.

The general principle is that data are collected in as detailed a way as possible using questions or sets of questions which are likely to yield the most accurate answers. For analysis purposes however, combining variables, reducing the number of categories etc. may be required. This kind of variable construction should never be done by the interviewer, nor at the manual editing/coding stage, as errors are inevitably introduced. Moreover it would limit subsequent flexibility in the way new variables can be constructed.

Although much data analysis software allows for variable construction and recoding at the time of analysis, it is useful to have variables that are going to be used repeatedly for analytical purposes available on a file. In addition many existing data analysis packages require that the same amount of data exist for each respondent, i.e. that the data have been rectangularized. It is therefore useful to create a new file which:

- Contains one record of fixed format per respondent.
- Contains all the variables that are most likely to be used for analysis. These may include original questions as well as constructed variables.

5.2 The WFS Standard Recode File

A recoded file for the WFS data, known as the "standard recode" file, is created for three purposes:

- To simplify the production of Country Report No. 1 tables.
- To provide a general 'analysis' file for researchers wanting to do further analysis on the data.

- To provide a standard set of variables for all countries participating in the WFS making comparative analysis possible.

The exact contents of a WFS Standard Recode file together with the rules by which each variable is constructed are given in the recode specifications in Appendix II.

The variables in this file are in two distinct sets. The first set is derived mainly from the birth and marriage histories. This set is generated automatically if the special WFS date edit, imputation, and recoding program (DEIR[1]) is used. The second set of variables does not depend on the first (except for a few key variables which can easily be recreated from the questionnaire) and can be generated independently.

The resulting file is the basis for the Standard Recode Tape which is documented and archived at the WFS headquarters and in the relevant country. The tape that is archived normally contains the data file, a codebook in the form of a machine-readable WFS dictionary, the marginals (ready for printing) of all variables and the SPSS[2] variable description for the data (derived from the WFS dictionary). The data are then in an easy-to-use form for further analysis.

5.3 Recoding Conventions

The recoded data file in general contains one fixed length, fixed format record per respondent. This ensures that analysis is possible with the data analysis software available at many computer installations. Conventions on the contents of each data record in WFS standard recode files to facilitate the use of data are outlined below.

5.3.1 Grouped Variables

Certain groupings of numerical variables (e.g. age) may frequently be required for tabulation purposes. In this case it is desirable to create new variables accordingly (e.g. age in 5-year groups, age in 10-year groups). When this is done, the original ungrouped variable is always retained as well in the standard recode file. This makes possible different ad hoc groupings in the future.

5.3.2. Leading zeros

The contents of each data record are numeric (again because much data analysis software requires this). In addition, where a field is longer than necessary to accommodate a value, leading zeros are inserted, not blanks. For example a value of 7 in a two digit field will be recorded as 07.

5.3.3 Missing data

Some computer software does not distinguish between blanks and zeros. Other software will not accept blanks at all. Blanks should therefore never be left in the data when information is

[1]DEIR Users Manual WFS Tech. 1430.
[2]Statistical Package for the Social Sciences.

missing. Special codes are assigned for missing data distinguishing, if desired, between different types with different codes. In WFS data, fields of 8's are used to indicate "not applicable" responses and fields of 9's indicate "not stated". Care must be taken that there is no valid response equivalent to the missing data code. For example for a variable "husband's age", the value 88 could mean either an age of 88 or not applicable — woman not married. In such a case either all ages over 87 should be recoded with code 87 meaning 87+ *or* the field should be expanded to three or four digits where 88 will be an age and 888 or 8888 will mean not applicable.

5.3.4 Default codes

When programming recode instructions some special value should be assigned to each variable in case no recode instruction defines a value. Presence of this value in the output record can then be taken as evidence of a gap in the recode program or an error/inconsistency in the input data. It is therefore recommended that at the beginning of the program for constructing the recoded data record, the entire data record be first filled with 7's i.e. 77 for a 2 digit or 7777 for a 4 digit field. The final version of the recode file should not contain any of these default codes.

5.3.5 Respondent Identification

It is important that each record in the recode file contain a unique identification. This enables reference to individual records for correction or listing purposes. During data cleaning stages this identification must be the same as on the questionnaire so that data records with errors can be referred back to their original source document (the questionnaire). Unless confidentiality considerations make this undesirable, it is recommended that the same questionnaire identification be retained on the recoded data record. Reference back to questionnaires may still be necessary when data are in doubt. Matching records on the recode file with other data about the same respondents (e.g. from the household schedule) is only possible if the same identification exists on the records in each data file.

5.4 Recode Specifications

Appendix II gives complete specifications of all variables in the WFS Standard Recode data record. The recode instructions are expressed in terms of questions from the WFS Core Questionnaires and modifications and assume the coding scheme described in the WFS *editing and coding manual* or in the codebook also given in Appendix II.

5.4.1 Standard Recode File Contents

The variables can be divided into two sets. The first set (V001-V306) consists of identification variables followed by variables dealing with age, nuptiality and fertility. Data records containing these variables are generated automatically if the WFS program DEIR is used for date edit, imputation, and recoding. The recode specifications for these variables given in Appendix II are for when DEIR is not used and assume that a fully edited and date imputed questionnaire file is available in which all dates are in the month/year form.

The second set (V401-V907) consists of variables dealing with exposure status, fertility preferences, contraception and background characteristics. The recode specifications make no

reference to dates in birth and marriage histories nor to recode variables from the first set (except for five key variables which can easily be recreated from the questionnaire). Thus the process of creating these variables requires only an edited questionnaire file as described in Chapter 3 and can proceed in parallel with the editing and imputation of dates.

Additional, country specific variables are constructed as required and added after V907 of each record.

The parts of the file can be merged to obtain the complete recode file.

5.4.2 Variables Names

The recoded variables have been classified into 10 major substantive groups. Each variable has a name consisting of the letter V followed by three digits. The first digit identifies the group (0-9) and the other two form a serial number. The only exceptions are the birth and marriage histories where the names start with B and M respectively, followed by 3 digits e.g. B034 represents the fourth item of the third birth. Country specific variables are given names starting with S. As far as possible these are grouped into sections reflecting the sections in the standard variables.

5.4.3 Code ranges

For most variables, minimum and maximum codes are given in the recode specifications. These minimum and maximum codes are the logical possible values as defined by the questionnaire and the editing and coding instructions. For example, if the questionnaire only allows for the current marriage plus four others then the maximum for the variable "number of times married" (V101) is 5. Similarly if it is an ever married sample, the minimum code for this variable will be 1.

In a few variables where the logical range is not obvious, an "*" is given instead of minimum and maximum codes. In these cases, the final dictionary for the data will show the actual minimum and maximum codes as given in the marginals for those variables.

5.4.4 Diagrammatic representation of recode specifications

A recode variable's value is set according to the values of other already existing variables (i.e. questions from the original questionnaire or previously constructed recode variables).

Some recoding involves a simple arithmetic computation e.g. to construct the "century month code" for the respondent's data of birth (V008) from the year (Q107Y) and month (Q107M) of birth:

$$V008 = 12 * Q107Y + Q107M$$

(We use the convention that Vn represents a recode variable and Qn represents a question number from the original questionnaire).

The rules for other types of recoding can generally be expressed in the following form:

If C_1 then r_1
else if C_2 then r_2
else if C_3 then r_3

where C_i is a condition
r_i is the value the recode variable will take when the condition C_i is true.

The conditions C_i may be simple, depending only on the values of a single variable, in which case the recode rules can be expressed in a diagrammatic way:

If	C_1	C_2	C_3 ----------
Then	r_1	r_2	r_3 ----------

e.g. for a 10 year grouping of age (recode variable V012) based on an existing variable for age (recode variable V010) we get

If V010 =	<25	25-34	35-44	45+
Then V012 =	1	2	3	4

For more complex conditions we can use the following diagrammatic representation:

<table>
<tr><th>If</th><th colspan="3"></th><th>Then</th></tr>
<tr><td colspan="3">C_1</td><td>r_1</td></tr>
<tr><td rowspan="2">C_2</td><td colspan="2">C_3</td><td>r_2</td></tr>
<tr><td colspan="2">C_4</td><td>r_3</td></tr>
<tr><td>C_5</td><td colspan="2">C_6</td><td>r_4</td></tr>
<tr><td rowspan="2">C_5</td><td>C_7</td><td>C_8</td><td>r_5</td></tr>
<tr><td></td><td>C_9</td><td>r_6</td></tr>
</table>

The vertical lines between conditions represent "and". All conditions along a row must be true to give the indicated recode value in that row. The diagram is equivalent to the statements

If C_1 then r_1
else if C_2 and C_3 then r_2
else if C_2 and C_4 then r_3
else if C_5 and C_6 then r_4 etc.

An example of this type of recoding is given by the rules for constructing the "pattern of work" standard recode variable (V713) based on questions Q601 (whether currently working), Q602 (ever worked since first marriage) and Q613 (worked before first marriage)

If			Then
Q601 = 9			99
Q601 = 1	Q613 = 1		1
	Q613 ≠ 1		2
Q601 = 2	Q602 = 1	Q613 = 1	3
		Q613 ≠ 1	4
	Q602 ≠ 1	Q613 = 1	5
		Q613 ≠ 1	6

The recoding specifications for the variables in the standard recode file as given in Appendix II are mostly expressed in one of the above ways.

6.1 Tabulations

6.1 Introduction

One of the principal methods of primary analysis of survey data is through tabulation. More sophisticated multivariate analysis of the data may follow once basic tables are available. Thus, for the WFS, tabulations form a major part of the First Country Report. The recommended tables for this report are given in *Guidelines for Country Report No. 1.* Detailed specifications of these tables will be found in Appendix II of this document. This chapter discusses some general principles of table building and introduces some terminology.

6.2 Software for Table Production

Many software packages exist for making tabulations of data and one such package should be used. These packages however vary greatly in their hardware requirements, their ability to construct new variables at the time of table building and the flexibility of their printout format and labelling options. Having installed such a package, programming for tabulations is a matter of setting up commands in the package's control language to produce the desired tables.

It is convenient if the final tables for a report can be produced from the computer in "camera ready" form, i.e. a form suitable for direct photo-reproduction and publication. This saves all the errors that are inevitably introduced if tables are typed. This of course requires not only that suitable software be available but also that all required table titles and labels be entered into the computer.

The first versions of tables are not always the same as the final versions decided upon for publication. It is therefore not worth paying too much attention to neat labelling and formatting at this stage although some basic table identification (e.g. a table number) and row and column headings should be introduced for analyst readability. After the first versions of tables have been examined, survey staff can make any necessary amendments, modifications, deletions etc. and decide on the final set of tables and their format for inclusion in the report. This final set should then be produced, wherever possible, with a package that prints with comprehensive labelling and titling in a camera ready form e.g. COCENTS[1].

It should be noted that for the set of standard tables recommended by the WFS for the First Country Report, all titles, footnotes and row and column headings are available in machine readable form on tape. Minor modifications can be made to these for country specific labelling and variable groupings and they can be used in conjunction with COCGEN, the WFS preprocessor for COCENTS, to produce fully labelled tables directly. Initial versions of tables can also be produced in the camera ready format with no extra work in this case.

[1]COCENTS. US Bureau of Census Tabulation Package.

6.3 Specifications of Tables

Any table can be specified in terms of:

- The population to be included.
- The classification variables defining panels, rows, and columns.
- The cell entries.

Although tabulation software sometimes provides for recoding of variables while the data are being read, the prior production of a "recode" file (as outlined in Chapter 5) containing all variables in exactly the form required for the tables simplifies the preparation of control commands for table production. Thus, for example, the WFS Standard Recode data records contain all variables that are required for the production of the tables recommended for WFS First Country Reports.

Given that all variables have been recoded suitably, then generally each required table can be specified in terms of up to five parameters giving, respectively, the base population, the rows, the columns, the panels, and the cell entries.

6.3.1 Variables Defining the Base Population

Tables which are to be restricted to a particular sub-population can be specified in the terms of a "base" variable such that only records with the base variable coded a particular value will be included in the table. When using the COCENTS tabulation package, base variables are most easily used if they are coded either 0 or 1 where cases coded 1 are to be included and cases coded 0 are to be excluded. For this reason the WFS standard recode data contains a set of variables coded 0, 1 which are used as base variables in Country Report No. 1 tables. Cases may appear with NA or NS codes on such base variables and need also to be excluded from the table (by treating such codes as 0's).

The base for a table is selected so that once it is confined to cases with the base variable = 1, cases with NA codes for any other variables involved in the table are automatically excluded. For example, in the WFS Country Report No. 1, Table 5.2.5 is restricted to a base of women who are currently pregnant or have had at least one live birth (V234 = 1). The column variables (V640 and V641) have been coded as NA for women who are not currently pregnant and have not had a live birth (i.e. which have V234 ≠ 1). These cases are therefore automatically excluded from the table once cases with V234 ≠ 1 are excluded.

In the examples of types of tables below, when the base variable is specified as "ALL" this means that the population to be included in the tables is unrestricted, i.e. the table covers all the cases in the file; "EM" implies tables covering all ever-married women.

6.3.2 Classification Variables (Rows, Columns, Panels)

The cell of a table to which a particular case belongs is defined by the values of classification variables for that case. If there are only two classification variables, then their categories define the row and column of a two dimensional table. When a third classification variable is introduced we call it the panel variable where each category can be thought of as defining a

6.1 Tabulations

6.1 Introduction

One of the principal methods of primary analysis of survey data is through tabulation. More sophisticated multivariate analysis of the data may follow once basic tables are available. Thus, for the WFS, tabulations form a major part of the First Country Report. The recommended tables for this report are given in *Guidelines for Country Report No. 1.* Detailed specifications of these tables will be found in Appendix II of this document. This chapter discusses some general principles of table building and introduces some terminology.

6.2 Software for Table Production

Many software packages exist for making tabulations of data and one such package should be used. These packages however vary greatly in their hardware requirements, their ability to construct new variables at the time of table building and the flexibility of their printout format and labelling options. Having installed such a package, programming for tabulations is a matter of setting up commands in the package's control language to produce the desired tables.

It is convenient if the final tables for a report can be produced from the computer in "camera ready" form, i.e. a form suitable for direct photo-reproduction and publication. This saves all the errors that are inevitably introduced if tables are typed. This of course requires not only that suitable software be available but also that all required table titles and labels be entered into the computer.

The first versions of tables are not always the same as the final versions decided upon for publication. It is therefore not worth paying too much attention to neat labelling and formatting at this stage although some basic table identification (e.g. a table number) and row and column headings should be introduced for analyst readability. After the first versions of tables have been examined, survey staff can make any necessary amendments, modifications, deletions etc. and decide on the final set of tables and their format for inclusion in the report. This final set should then be produced, wherever possible, with a package that prints with comprehensive labelling and titling in a camera ready form e.g. COCENTS[1].

It should be noted that for the set of standard tables recommended by the WFS for the First Country Report, all titles, footnotes and row and column headings are available in machine readable form on tape. Minor modifications can be made to these for country specific labelling and variable groupings and they can be used in conjunction with COCGEN, the WFS preprocessor for COCENTS, to produce fully labelled tables directly. Initial versions of tables can also be produced in the camera ready format with no extra work in this case.

[1]COCENTS. US Bureau of Census Tabulation Package.

6.3 Specifications of Tables

Any table can be specified in terms of:

- The population to be included.
- The classification variables defining panels, rows, and columns.
- The cell entries.

Although tabulation software sometimes provides for recoding of variables while the data are being read, the prior production of a "recode" file (as outlined in Chapter 5) containing all variables in exactly the form required for the tables simplifies the preparation of control commands for table production. Thus, for example, the WFS Standard Recode data records contain all variables that are required for the production of the tables recommended for WFS First Country Reports.

Given that all variables have been recoded suitably, then generally each required table can be specified in terms of up to five parameters giving, respectively, the base population, the rows, the columns, the panels, and the cell entries.

6.3.1 Variables Defining the Base Population

Tables which are to be restricted to a particular sub-population can be specified in the terms of a "base" variable such that only records with the base variable coded a particular value will be included in the table. When using the COCENTS tabulation package, base variables are most easily used if they are coded either 0 or 1 where cases coded 1 are to be included and cases coded 0 are to be excluded. For this reason the WFS standard recode data contains a set of variables coded 0, 1 which are used as base variables in Country Report No. 1 tables. Cases may appear with NA or NS codes on such base variables and need also to be excluded from the table (by treating such codes as 0's).

The base for a table is selected so that once it is confined to cases with the base variable = 1, cases with NA codes for any other variables involved in the table are automatically excluded. For example, in the WFS Country Report No. 1, Table 5.2.5 is restricted to a base of women who are currently pregnant or have had at least one live birth (V234 = 1). The column variables (V640 and V641) have been coded as NA for women who are not currently pregnant and have not had a live birth (i.e. which have V234 ≠ 1). These cases are therefore automatically excluded from the table once cases with V234 ≠ 1 are excluded.

In the examples of types of tables below, when the base variable is specified as "ALL" this means that the population to be included in the tables is unrestricted, i.e. the table covers all the cases in the file; "EM" implies tables covering all ever-married women.

6.3.2 Classification Variables (Rows, Columns, Panels)

The cell of a table to which a particular case belongs is defined by the values of classification variables for that case. If there are only two classification variables, then their categories define the row and column of a two dimensional table. When a third classification variable is introduced we call it the panel variable where each category can be thought of as defining a

slice or panel of a three dimensional figure. N dimensional tables can always be reduced to a three dimensional table with a single panel variable provided a recode variable for the panel is constructed which has categories for each possible combination of the additional classification variables.

It is convenient if classification variables have been recoded in such a way that the categories involved are numbered sequentially, starting with "1". For example, in WFS Country Report No. 1, Table 1.5.1, the two panels correspond to V112 = 1,2. Categories for certain other variables may start with "0" and increase consecutively from that.

6.3.3 Variables Defining Cell Entries

The values of these variables are used to define means or proportions appearing in cells of a table. Further explanation will be given in Section 6.4 where six different types of tables are considered.

6.3.4 The Presentation of "DK" or "NS" Categories

As mentioned earlier, cases with NA codes on any of the variables being used to define the parameters of a table are always eliminated from the table, i.e. the NA code never appears in any table. However, the NS code, where present, must always appear in the table in the form of an extra panel, row or column as the case may be. The exact form of presentation of the NS cases depends on the nature of the cell entries in the tables.

In accordance with the WFS *Editing and Coding Manual,* generally no distinction has been made between "DK" (the respondent does not know the answer) and "NS" (no answer has been recorded by the interviewer).

In general, tables which have NS cases on the row and/or column variables are presented with two extra rows and/or columns showing subtotals for cases with stated values and distributions of the not stated cases. Such tables look as follows:

	Column categories				ST	NS	Tot
Row categories							
Subtotal							
Not stated							
Total							

For tables with a cell variable, (i.e. tables of means, proportions or percentages) cases which have NS values for the cell variable are never included in the main body of the table. However, an extra row and column of the frequencies of such cases is given (see sample table type iv Section 6.4.7 below).

6.4 Types of Tables

The six types of tables described below are those used in WFS Country Reports No. 1.

6.4.1 Table Giving Simple Frequencies (FREQ[1])

In the tables, all codes of the classificatory variables, with the obvious exception of the NA code, must appear. In particular, the NS code should appear as an extra panel, row or column, as appropriate. Marginal frequencies, giving sums for all rows, all columns and the whole table, are always included.

6.4.2 Tables Giving Row Percentages for Categorical Variables (CATG[1])

In these tables the column variable consists of simple categories like those of marital status, exposure status, pattern of contraceptive use, etc. Here cell frequencies have been divided by row totals to give percentage figures. A marginal column giving the row total frequencies (= 100 per cent) is also included. Similarly, a marginal row showing the percentage figures for each column should also be given.

6.4.3 Tables Giving Row Percentages for Metric Variables (FREQ, METR[1])

In these tables the column variable has an interval level of measurement. Examples are variables like number of living children, length of the open birth interval, etc. The layout of these tables is similar to the types FRMN and CATG described above, except that another column, giving the row mean for the column variable, is added (a ratio column may also be added — see example (III) below).

6.4.4 Tables Giving Cell-by-Cell Percent (PERC[1])

A percentage is formed by multiplying by 100 the number of cases having a particular attribute in a given cell, then dividing the result by the total number of cases belonging to that cell. An example is the percentage of currently married women (the cases) who are pregnant (the attribute). A cell variable defines the attribute as YES or NO. All cases with NS for the cell variable are excluded from both the numerator and the denominator when computing the percentage. However, such cases are not entirely excluded: they are counted and included as a final row and a column outside of the main table.

[1]The four digit keywords are those used in the WFS Program COCGEN (a preprocessor for the COCENTS tabulation package) for specifying the type of table.

In constructing recode variables that are going to be used as cell variables for percentages, the following system is used in the WFS Standard Recode:

YES is coded as 01
NO is coded as 00
NS is coded as 99
NA is coded as 88

Of these four codes, NA is automatically excluded from the table by selecting an appropriate base variable.

6.4.5 Tables Giving Cell-by-Cell Means (MEAN[1])

The layout for these tables is identical to that for the previous type. The denominator for the mean is again the number of cases belonging to a cell. The numerator is the sum of values of the cell variable for the same cases.

6.4.6 Tables Giving Cell-by-Cell Ratios (RATI[1])

The cell entries for these tables are defined by the ratio of two cell variables, Vn and Vd. The numerator consists of the sum of values of Vn for the N cases belonging to the cell. The denominator consists of the sum of values of Vd for the same N cases. Three figures are given for each cell:

- The ratio expressed as a percentage value $= \dfrac{\text{Sum of values of Vn for N cases}}{\text{Sum of values of Vd for N cases}} \times 100$
- The mean of the denominator variable $= \dfrac{\text{Sum of values of Vd for N cases}}{N}$
- The base frequency $= N$

6.4.7 Examples of the Types of Tables

On the following pages are examples of all types of tables described above.

(i) FREQ table (Table 1.1.2)
(ii) CATG table (Table 1.5.1)
(iii) METR table (with a ratio column) (Table 2.3.2)
(iv) PERC table (Table 3.1.3.A)
(v) MEAN table (Table 2.2.7b)
(vi) RATI table (Table 2.4.4)

Note: In all these examples, totals are shown in the first column instead of the more usual last column position.

6.5 Tables For Weighted Data

Substantive results are always presented for weighted data (if applicable) with weighted frequencies. If the unweighted frequencies differ substantially (say by more than 30 percent)

[1]The four digit keywords are those used in the WFS Program COCGEN (a preprocessor for the COCENTS tabulation package) for specifying the type of table.

from the weighted frequencies, then these should also be given. This may be done in each individual cell, or as a separate panel of unweighted frequencies accompanying each table.

If limitations in the available software make it difficult to construct tables showing both the weighted and unweighted frequencies simultaneously then a limited number of tables giving unweighted frequencies which provide the required information (or approximations) for a majority of the main tabulations should be shown in a separate appendix.

In addition where main tables are of various dependent variables by a set of different background variables, then unweighted tables should be made of the various background variables against one another. The total number of additional tables can be quite large though the tables involved are generally much simpler than the main tables.

The question of weighting of sample data, including that of presentation of weighted results, is discussed in *Guidelines for the Country Report No. 1,* Appendix V.

6.6 Sampling Errors For WFS Country Report Tabulations

Sampling errors for the main survey estimates should be included along with the detailed cross-tabulations in the First Country Report. Appendix II outlines a recommended set of variables and sub-classes that may be derived from the standard recode tape and gives a brief exposition on how those are specified for the CLUSTERS program developed by the WFS for computing sampling errors on clustered samples. For details the Users' Manual for CLUSTERS[1] should be consulted.

[1] Users' Manual for CLUSTERS WFS/TECH 770 June 1978.

Table 1.1.2 (1st page)

(i) Example of FREQ Table

Distribution Of All Women Ever In A Union According To Calendar Year Of Birth — By Age At First Union In Single Years

AGE AT INITIAL UNION	AT TOTAL	YEAR OF BIRTH 1925	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935
TOTAL	3,616	19	58	72	78	74	112	72	77	93	70	106
8—10	11	—	—	—	—	—	—	—	—	—	—	—
11	28	—	—	—	5	1	1	1	2	—	1	—
12	57	—	2	3	—	4	1	2	3	1	—	2
13	141	3	7	5	1	—	3	2	5	4	3	3
14	308	—	4	5	6	6	10	3	3	12	4	12
15	491	—	4	12	11	6	16	8	12	15	5	13
16	562	6	1	9	7	8	12	14	14	16	14	18
17	496	3	3	15	7	11	22	10	10	10	8	13
18	402	—	7	5	9	6	13	6	4	6	9	9
19	336	2	9	7	4	3	8	7	5	8	4	5
20	229	1	8	2	4	11	8	1	5	2	7	7
21	174	2	3	3	4	1	3	8	6	5	4	6
22	110	1	4	1	4	—	3	4	2	2	1	4
23	93	—	2	1	5	3	3	3	1	1	6	3
24	48	1	—	—	1	2	—	1	1	4	—	4
25	35	—	—	—	2	3	2	—	—	2	—	2
26	22	—	1	1	1	—	1	—	1	2	1	1
27	19	—	1	—	2	—	4	—	—	—	1	—
28	15	—	—	—	2	4	—	—	—	—	—	1
29	6	—	—	—	—	2	—	—	1	—	1	1
30	4	—	—	1	—	1	—	—	—	—	—	—
31	7	—	—	—	2	1	—	—	—	—	—	1
32	6	—	—	1	—	—	1	—	1	—	—	1
33	3	—	—	—	—	—	—	—	—	1	—	-
34	3	—	1	—	—	—	—	1	—	—	—	—
35	5	—	—	1	—	1	1	—	1	1	—	—
36	3	—	—	—	1	—	—	—	—	—	1	—
37	—	—	—	—	—	—	—	—	—	—	—	—
38	1	—	—	—	—	—	—	1	—	—	—	—
39	—	—	—	—	—	—	—	—	—	—	—	—
40	—	—	—	—	—	—	—	—	—	—	—	—
41	—	—	—	—	—	—	—	—	—	—	—	—
42	—	—	—	—	—	—	—	—	—	—	—	—
43	—	—	—	—	—	—	—	—	—	—	—	—
44	—	—	—	—	—	—	—	—	—	—	—	—
45	—	—	—	—	—	—	—	—	—	—	—	—
46	1	—	1	—	—	—	—	—	—	—	—	—
47	—	—	—	—	—	—	—	—	—	—	—	
48	—	—	—	—	—	—	—	—	—	—	—	—
49	—	—	—	—	—	—	—	—	—	—	—	—

Table 1.5.1A (1st page) (ii) Example of CATG Table

The Percent Distribution Of All Women Ever In A Union According To Current Union Status — By Current Age And Level Of Education

LEVEL OF EDUCATION AND CURRENT AGE	CURRENT UNION STATUS				
	TOTAL	MARRIED	COMMON LAW	VISITING	SINGLE
ALL WOMEN EVER IN A UNION					
TOTAL	3,586	63.7	12.5	13.1	10.7
15—19	352	48.9	11.6	31.3	8.2
20—24	719	59.7	10.4	21.6	8.3
25—29	702	68.4	12.1	10.7	8.8
30—34	539	69.9	13.2	8.3	8.5
35—39	483	67.3	15.1	6.8	10.8
40—44	411	64.7	14.4	7.1	13.9
45—49	380	61.6	11.6	6.1	20.8
<4 YEARS PRIMARY					
TOTAL	593	71.2	15.5	2.4	11.0
15—19	22	54.5	31.8	9.1	4.5
20—24	52	69.2	13.5	9.6	7.7
25—29	81	72.8	22.2	1.2	3.7
30—34	102	81.4	11.8	3.9	2.9
35—39	112	70.5	19.6	.9	8.9
40—44	110	71.8	13.6	.9	13.6
45—49	114	64.9	9.6	.0	25.4
4+ YEARS PRIMARY					
TOTAL	1,694	64.8	14.6	9.2	11.3
15—19	47	72.3	8.5	19.1	.0
20—24	139	60.4	16.5	12.9	10.1
25—29	350	68.6	12.9	9.7	8.9
30—34	361	67.3	15.8	8.3	8.6
35—39	315	66.7	14.9	7.9	10.5
40—44	250	60.0	16.4	8.4	15.2
45—49	232	59.1	13.4	8.2	19.4
SECONDARY/HIGHER					
TOTAL	1,299	58.7	8.3	23.1	9.9
15—19	283	44.5	10.6	35.0	9.9
20—24	528	58.5	8.5	25.0	8.0
25—29	271	66.8	8.1	14.8	10.3
30—34	76	67.1	2.6	14.5	15.8
35—39	56	64.3	7.1	12.5	16.1
40—44	51	72.5	5.9	13.7	7.8
45—49	34	67.6	5.9	11.8	14.7

Table 2.3.2 (1st page) (iii) Example of METR Table (with ratio column)

The Percent Distribution Of All Women Ever In A Union According To The Number Of Living Children — By Years Since Entry Into Initial Union And Current Union Status

CURRENT UNION STATUS AND YEARS SINCE INITIAL UNION	TOTAL	NUMBER OF LIVING CHILDREN										MEAN NO. OF LIVING CHILDREN	PERCENT MALE
		0	1	2	3	4	5	6	7	8	9+		
ALL CURRENT UNION TYPES													
TOTAL	3,616	13.8	15.7	13.8	12.3	10.8	8.8	7.6	5.8	4.1	7.4	3.62	50.6
UNDER 5 YEARS	797	40.2	38.4	17.1	3.9	.3	.1	.0	.0	.0	.1	.91	52.7
5—9 YEARS	755	9.3	18.3	27.5	23.4	15.2	4.9	1.3	.0	.0	.0	2.46	51.3
10—14 YEARS	558	7.0	8.1	12.5	19.0	20.3	18.3	9.0	3.9	1.1	.9	3.63	50.8
15—19 YEARS	527	4.4	5.7	7.0	10.4	15.0	14.6	14.8	11.6	8.2	8.3	5.07	49.5
20—24 YEARS	429	5.1	4.2	4.0	8.6	8.6	12.4	16.1	14.0	10.3	16.8	5.80	49.7
25—29 YEARS	365	3.3	5.5	6.0	7.1	7.7	9.6	13.4	10.7	12.1	24.7	6.31	50.9
30 YEARS AND OVER	185	6.5	5.4	4.3	5.9	9.7	7.0	10.3	14.1	7.0	29.7	6.22	51.4
CURRENTLY MARRIED													
TOTAL	2,302	9.6	12.8	14.2	13.0	12.0	10.2	9.0	6.6	4.6	7.9	4.07	50.2
UNDER 5 YEARS	465	31.4	39.6	22.6	6.0	.4	.0	.0	.0	.0	.0	1.08	52.1
5—9 YEARS	457	5.9	9.6	30.6	27.6	18.2	6.1	2.0	.0	.0	.0	2.72	51.4
10—14 YEARS	384	5.2	5.5	10.2	19.0	23.4	20.1	10.7	3.9	1.3	.8	3.92	49.6
15—19 YEARS	369	1.9	4.9	6.0	7.9	16.0	14.9	16.0	13.3	9.5	9.8	5.43	50.0
20—24 YEARS	286	3.1	4.5	2.8	8.4	7.0	13.6	16.8	14.7	11.5	17.5	6.01	49.3
25—29 YEARS	235	2.6	3.4	4.7	6.8	6.8	11.9	15.3	12.3	11.1	25.1	6.54	50.4
30 YEARS AND OVER	106	5.7	5.7	.9	3.8	6.6	7.5	14.2	16.0	7.5	32.1	6.70	50.9
CURRENTLY COMMON LAW													
TOTAL	449	12.0	14.0	12.7	13.8	10.7	8.9	7.3	6.5	5.3	8.7	3.95	53.0
UNDER 5 YEARS	60	40.0	38.3	16.7	3.3	.0	1.7	.0	.0	.0	.0	.91	53.7
5—9 YEARS	105	6.7	23.8	23.8	20.0	18.1	6.7	1.0	.0	.0	.0	2.42	52.5
10—14 YEARS	68	7.4	7.4	11.8	20.6	14.7	17.6	10.3	7.4	1.5	1.5	3.84	59.1
15—19 YEARS	73	9.6	4.1	5.5	16.4	12.3	17.8	12.3	11.0	6.8	4.1	4.52	51.8
20—24 YEARS	72	9.7	1.4	6.9	9.7	9.7	5.6	13.9	12.5	9.7	20.8	5.66	51.4
25—29 YEARS	48	2.1	10.4	2.1	8.3	4.2	4.2	12.5	8.3	18.8	29.2	6.72	52.5
30 YEARS AND OVER	23	13.0	4.3	17.4	8.7	4.3	4.3	.0	13.0	8.7	26.1	5.20	50.0

TABLE 3.1.3A (1st page)

(iv) Example of PERC Table

The Percentage Of Woman Currently In A Union And "Fecund" Who Want No More Children — By Number Of Living Children (Including Any Current Pregnancy), Level Of Education And Current Age

CURRENT AGE AND LEVEL OF EDUCATION	TOTAL	NUMBER OF LIVING CHILDREN									
		0	1	2	3	4	5	6	7	8	9+
ALL AGES											
TOTAL											
NUMBER	3,041	311	476	458	413	346	291	228	173	126	219
PERCENT	51.1	7.7	15.8	35.8	51.8	59.5	78.0	84.6	84.4	86.5	89.5
<4 YEARS PRIMARY											
NUMBER	485	16	36	53	47	56	59	62	47	33	76
PERCENT	73.8	37.5	25.0	49.1	61.7	64.3	86.4	91.9	87.2	93.9	94.7
4+ YEARS PRIMARY											
NUMBER	1,393	79	118	144	210	192	178	149	116	78	129
PERCENT	61.3	13.9	20.3	35.4	54.3	62.5	77.0	81.9	83.6	84.6	86.8
SECONDARY/HIGHER											
NUMBER	1,146	214	321	259	155	95	53	14	9	14	12
PERCENT	51.0	2.8	13.1	33.2	45.2	49.5	71.7	78.6	77.8	85.7	83.3
NOT STATED											
NUMBER	17	2	1	2	1	3	1	3	1	1	2
PERCENT	76.5	50.0	.0	50.0	100.0	100.0	100.0	100.0	100.0	0.0	100.0
SUB-TOTAL EXCL. NOT STATED											
NUMBER	3,024	309	475	456	412	343	290	225	172	125	217
PERCENT	51.0	7.4	15.8	35.7	51.7	59.2	77.9	84.4	84.3	87.2	89.4

Table 2.2.7B (1st page) (v) Example of MEAN Table

Mean Number Of Children Ever Born To All Women
Ever In A Union — By Level Of Education,
Religion And Years Since Entry Into Initial Union

YEARS SINCE INITIAL UNION AND RELIGION	TOTAL	LEVEL OF EDUCATION				SUB-TOTAL EXCLUDING NOT STATED
		4 YEARS PRIMARY	4+ YEARS PRIMARY	SECONDARY/ HIGHER	NOT STATED	
ALL YEARS SINCE INITIAL UNION						
TOTAL						
NUMBER	3,616	593	1,694	1,299	30	3,586
MEAN	4.06	5.73	4.96	2.02	5.28	4.05
ROMAN CATHOLIC						
NUMBER	447	27	198	221	1	446
MEAN	3.22	4.71	4.41	1.93	1.00	3.22
ANGLICAN						
NUMBER	576	20	305	250	1	575
MEAN	3.71	4.81	5.12	2.01	6.00	3.71
HINDU						
NUMBER	1,302	417	551	314	20	1,282
MEAN	4.54	5.93	4.86	1.83	5.65	4.53
MUSLIM						
NUMBER	375	77	190	104	4	371
MEAN	4.13	5.44	4.61	2.12	4.00	4.13
OTHER						
NUMBER	916	52	450	410	4	912
MEAN	3.87	5.36	5.24	2.13	5.00	3.88

Table 2.4.4

(vi) Example of RATI Table

The Percentage Of Women Currently In A Union Reporting A Current Pregnancy — By Number Of Living Children And Current Age

CURRENT AGE	TOTAL	NUMBER OF LIVING CHILDREN									
		0	1	2	3	4	5	6	7	8	9+
TOTAL											
NUMBER	3,221	427	480	442	408	355	294	252	191	137	235
PERCENT	12.1	24.6	19.0	13.3	9.6	9.0	6.5	5.6	6.8	5.1	4.7
15—19											
NUMBER	326	140	141	41	3	1	—	—	—	—	—
PERCENT	26.1	32.1	26.2	4.9	.0	100.0	.0	.0	.0	.0	.0
20—24											
NUMBER	661	144	174	165	104	55	15	3	1	—	—
PERCENT	20.9	30.6	24.1	17.6	11.5	18.2	6.7	.0	.0	.0	.0
25—29											
NUMBER	642	66	76	135	134	108	79	33	6	2	3
PERCENT	14.5	19.7	13.2	16.3	14.9	12.0	13.9	9.1	16.7	.0	.0
30—34											
NUMBER	497	18	25	41	68	89	76	78	45	25	32
PERCENT	8.5	11.1	4.0	9.8	5.9	5.6	6.6	7.7	15.6	12.0	15.6
35—39											
NUMBER	432	12	22	31	43	44	57	58	60	45	60
PERCENT	5.3	.0	4.5	3.2	7.0	6.8	3.5	8.6	5.0	4.4	5.0
40—44											
NUMBER	359	31	21	17	32	29	37	35	38	40	79
PERCENT	2.2	3.2	.0	5.9	.0	.0	.0	.0	5.3	5.0	2.5
45—49											
NUMBER	304	16	21	12	24	29	30	45	41	25	61
PERCENT	.3	.0	.0	.0	.0	.0	.0	.0	.0	.0	1.6

Chapter 7 — Data Archiving

7.1 Types of Data To Be Archived

During the processing of survey data, a large number of different files are created. Once the data have been cleaned and the recode or 'analysis' file constructed, then the different files should be reviewed and either discarded or else fully documented and retained for possible future use. In general one should consider keeping at least three versions of the data:

- The original, uncleaned, raw data
- The cleaned raw data
- The recoded data

With the WFS data, there are various different types of respondents upon which analysis might be required: household members, individual eligible women and children. The original raw data contain information on all these. For analysis purposes it is convenient to separate them. It is therefore recommended that the following files be retained:

(i) Original, uncleaned raw data

(ii) Cleaned individual data with no imputed dates

(iii) Cleaned household data

(iv) Individual recode file (standard recode with imputed dates)

(v) Household member file (one record per household member with general household information added to each record)

(vi) optionally a children's file (one record per child with relevant mother's information added to each record).

7.2 Data Documentation

Data files in machine readable form are useless without associated documentation. The various types of documentation considered necessary for data to be archived and to be usable by researchers for further analysis are outlined below.

7.2.1 Codebook

The basic documentation for the actual data is the codebook. As described in Section 1.2.4 this specifies each variable that is in a data record, giving its location in the record, its name, its "missing data" codes and a description of the meaning of each code. Machine readable versions of this information are extremely useful for those analysing the data. Thus WFS standard recode data files on tape are always accompanied by such a codebook known as the *WFS dictionary.*

7.2.2 Description of the Survey

The codebook in itself is not comprehensive enough to give a full understanding of the data.

A written document containing special notes about the survey and the way it was conducted is also essential. The following information is recommended as a basis for this document:

(i) A statement of the nature of the data being documented, a list of the different files of data available and references to related documents.

(ii) The name of the executing agency which carried out the survey.

(iii) A description of the sample. This should include whether it was stratified, the number of area stages and number of clusters and whether it is self-weighted. If weights are used, then indication should be made of whether they correct only for unequal final probabilities or also for differential non-response. The sum of weights should be given as well as the rules by which the weights are assigned to the different respondents in the data.

(iv) A short description of the questionnaire. For the WFS this might mention deviations from the Core Questionnaire and the extra modules used. For recode files, any section of the questionnaire *not* in the file should be mentioned.

(v) Details of the field work and office editing and coding giving the numbers of people involved and the dates each stage took place. In addition a short comment on editing procedures and a list of the edit checks used could be given.

(vi) Data processing methods and software used for checking, correcting, imputing, recoding and tabulating the data.

(vii) Imputation procedure and a summary of which variables and for how many cases imputation was done.

(viii) Any other information and peculiarities of the survey data collection and processing not noted before.

(ix) The structure of the data file: whether it is hierarchial or "rectangular"; if there is more than one record per case, details of the different card types and whether they are obligatory or optional; the way the file is sorted.

(x) explanatory notes on individual variables where there are for example known errors or deficiencies or where further explanation beyond that given in the codebook is required.

7.2.3. Questionnaire and coding instructions

A copy of the original questionnaire should always be available. Where codes are not given on the questionnaire itself, then the coding instructions used to code the data are also required.

7.2.4 Editing and Recode Specifications

All specifications developed for processing the questionnaires (as given in Appendix II) should be updated to reflect exactly what was done. This is especially important for the recode specifications. These are the basic definition of the variables in the recode file used for all subsequent data analysis and should reflect the actual recode program used to generate the file. The complete DP specifications comprise an important part of the documentation for the data.

7.2.5 Computer Programs Used

Copies of the special computer programs (or control commands used with general purpose programs) for achieving the data cleaning and recoding must be retained. Questions on particular data values may only be resolved by recourse to these.

7.2.6. The DP Manual

If a comprehensive DP manual is kept up to date throughout the survey data processing as recommended in Section 1.5.1, this itself provides complete documentation for the data including most of the above items.

7.3 Marginal Distributions

Any data analyst needs to know the distributions on the variables that are to be used for analysis before proceeding. Such distributions may be produced when required but it is convenient to have them archived with the data for quick and easy reference. These should be both the unweighted distributions and, if the data are weighted, the weighted distributions.
For all data in the WFS Archive, printout for the marginals is stored on tape along with the data and the codebook.

Appendix I

Some Software for Data Cleaning, Tabulation and Further Analysis

Appendix I

Some Software for Data Cleaning, Tabulation, and Further Analysis

Name	Purpose	Source	Language	Hardware & Minimum Core Requirement
I Data Editing and Recoding				
Utilities	Copying, listing, sorting	Normally available at computer installation		
CONCOR	Compiler and executor for a purpose built language for range and consistency checking and for recoding.	(i) CELADE, Santiago Chile (ii) WFS headquarters	IBM Assembler	IBM 360/370 OS/DOS 50K bytes
DEIR	Special purpose program for checking dates in the marriage and birth histories of WFS data. Optionally imputes missing months. Optionally creates an output file containing variables V001-V306 of the WFS standard recode file.	WFS headquarters	COBOL	IBM 360/370 OS/DOS HP 3000 Easily converted for other hardware 64K bytes
OSIRIS III OSIRIS IV	General purpose packages for data cleaning, data management and data analysis. OSIRIS IV handles structured files and is an interactively oriented system. Otherwise the same facilities are available in both. Good structure checking, range and consistency checking and file merging procedures.	Institute for Social Research, University of Michigan, Ann Arbor, MI 48106, USA	FORTRAN + ASSEMBLER	IBM 360/370 OS 106K bytes
PSTAT	General purpose package for data management and analysis. Good file matching and merging and some data editing procedures.	PSTAT Inc. P.O. Box 285 Princeton New Jersey, USA	FORTRAN	Many different computers including IBM 360/370, PDP 10. 256K bytes (Mini PSTAT approx. 150K)

Name	Purpose	Source	Language	Hardware & Minimum Core Requirement
STRUCT	Checks the structure of card image data against user supplied specifications of required and optional card types. Optionally produces a file containing only cases with correct structure with cards deleted/inserted according to options chosen.	WFS headquarters	FORTRAN	IBM 360/370 HP 3000 50K bytes
SUPDATE UPDATE	Delete, insert, replace, modify records in a card image file. SUPDATE uses the serial number of the record in the file while UPDATE uses an identification field to identify which records are to be updated. One column (normally 80) is used as a transaction code and must therefore *not* be used to accommodate information on the data cards.	WFS headquarters	COBOL	IBM 360/370 HP 3000 ICL 1900 50K bytes.

Note: SPSS is more usable than is often thought for editing and recoding, provided one has a rectangular file.

II Data Tabulation

CENTSAID	Generates COCENTS table, building statements from an easy to use language. Similar function to COCGEN although no library available for WFS tables.	Dualabs 1601 N. Kent St, Arlington, Virginia 22207 USA	COBOL	IBM 360/370 Honeywell 6000 96K bytes

Name	Purpose	Source	Language	Hardware & Minimum Core Requirement
COCGEN	Generates COCENTS table building statements from a more user oriented language. A library containing all COCGEN specifications necessary to produce WFS standard tables is also available.	WFS headquarters	COBOL	IBM 360/370 ICL 1900 HP 3000 64K bytes
COCENTS	Table building package. Can produce tables in a form suitable for direct photocopying. Control language difficult (see COCGEN).	U.S. Bureau of the Census Washington D.C. USA	COBOL	Most computers 32K bytes
CENTS	Same function as COCENTS and more efficient but only available for IBM 360/370.	U.S. Bureau of the Census	IBM 360/ Assembler	IBM 360/370 OS/DOS 32K bytes
MARG	Marginal distributions from structured files. Uses a "CONCOR" dictionary to define variables. Slow but sure.	WFS headquarters	COBOL	IBM 360/370 64K bytes
OSIRIS	See above. Also has table building capabilities, including marginals.	See above	See above	See above
PSTAT	See above. Has good table building capabilities, including marginals.	See above	See above	See above
SPSS	General data analysis. Very commonly available and very easy to use. Limited control of table formats in versions prior to version 8. Very easy to produce marginals.	SPSS Inc Chicago USA	FORTRAN	Many different computers, including IBM 360/370, HP 3000, ICL 1900/2900 200K bytes

Name	Purpose	Source	Language	Hardware & Minimum Core Requirement
TPL	Table Producing Language. Very powerful and relatively easy to use. Output suitable for direct photocopying.	US Dept. of Labour Washington D.C. USA	PL/1	IBM 360/370 OS 300Kbytes
III General Data Analysis				
CLUSTERS	Computes sampling errors for clustered samples.	WFS headquarters	FORTRAN	Most computers 64K bytes
FERTRATE	Program for computing different types of fertility rates from WFS standard files.	WFS headquarters	FORTRAN	HP 3000 easily converted 64K bytes
BMDP	A large package containing many different analysis programs including regression and life tables.	UCLA California USA	FORTRAN	Many different computers, including IBM 360/370 HP 3000
SPSS	General data analysis, including regression and life table analysis. (in Version 8)	See above	See above	See above

Note: Work is underway at the WFS headquarters to produce

(i) a data editing and data management package, including format and structure checking, updating, listing, and file merging capabilities.

(ii) a data analysis package for specialized techniques not normally found in large general purpose packages.

This software will have a common type of control language and will access data descriptive information from the WFS dictionary.