Chapter 5 - Data Processing

This chapter presents:

- *Data processing overview*
- *Data preparation, validation and correction*
- *Tabulation and other analysis*
- *Data processing calendar and organization*
- *Adapting the generic system to country requirements*

The key to producing fast, accurate and useful survey results is a well designed, tested and documented data processing system ready to process questionnaires as they arrive from the field. The CWIQ package includes a completely operational data processing system with programs designed to process the generic questionnaire. The system also includes operating instructions, system specifications, and technical documentation needed to adapt the system to country specific requirements. The generic system is used with minimum change to process the pilot survey; it is adapted to process the first national survey; and it can be extended with modules for the second and subsequent national surveys. This chapter describes the design and operation of the generic CWIQ data processing system and its adaptation to a country specific questionnaire.

Data Processing Overview

The CWIQ processing system resembles a typical survey data processing system with three main steps: data entry, validation and tabulation. However, the CWIQ system has some unique characteristics. First, the questionnaire has been designed to be processed using optical scanners, eliminating the need for keying data. Second, the system uses a relational database (Microsoft Access 97) to store the survey data instead of text or proprietary format files. Third, the system has been programmed in a general purpose programming language (Microsoft Visual Basic for Applications) instead of a generalized survey data processing package.

A principal objective of the CWIQ is to produce publication ready tables within one month of the end of fieldwork. To achieve this, data processing must begin as soon as completed questionnaires are available from the field. Data processing is done at the central survey office by a small team of clerks and machine operators under the supervision of the data processing coordinator. The data processing team is responsible for handling the survey questionnaires from the time they arrive from the field until they are stored in the survey archive. Data processing is carried out concurrently with fieldwork. Questionnaires from completed clusters are sent from the field to the central office for processing as soon they are available. Questionnaires are processed by cluster through the data preparation and validation stages of the system. The last clusters should be scanned and fully validated within two weeks of the end of fieldwork, and the standard tabulations should be available two weeks later.

For the pilot survey, the team is equipped with one or two computers, a scanner and a printer. Equipment requirements for the national survey will depend on the sample size and the survey calendar. Complete hardware and software specifications are in Annex 6.

Data Preparation, Validation and Correction.

Data Preparation

To insure a steady flow of questionnaires from the field, questionnaires for a cluster are reviewed for completeness and packaged for delivery to the survey office when all the interviews in the cluster have been completed.

When a cluster package arrives from the field it is logged in and prepared for input into the system. The questionnaires are counted and sorted by household identification. A manual edit of the cover page of the questionnaire and the reference number written on the top of all pages is done to detect errors that cause problems during and after scanning. The questionnaires and all accompanying documents are put into a file folder labeled with the cluster number. The date of arrival, the number of questionnaires and the person assigned to process the cluster is recorded in the master reception log, on the folder and updated in the master cluster database.

The data entry step of the system consists of “reading” the questionnaires using an optical scanner and storing their images on disk. The questionnaires are fed through the scanner rather like pages through a photocopier. During scanning, the scanner creates an image of each page of each questionnaire. The scanned images are subsequently evaluated by the scanning software and questionnaires with possible errors are subject to verification by the scanner operator. Typical errors include unidentified forms that can not be evaluated, questionnaires with missing or mismatched pages, unrecognizable hand printed characters and questionable filling of bubbles.

After verification, the cluster data are transferred via diskette or local area network to the central computer for processing. Further checks are performed at this stage to detect duplicate questionnaires and other serious errors which need to be corrected before proceeding. To insure completeness, the number of questionnaires scanned is compared to the number of households selected in the sample for that cluster. If there are any errors, the list of households scanned is compared to the questionnaires. All errors must be resolved before proceeding to the next step.

When there are no errors, the cluster data are transferred to the questionnaire database for validation. This database is used for subsequent validation, correction and tabulation. Note, that a copy of the original questionnaire data is retained as a backup and to permit analysis of the “unedited” data for data quality and methodological purposes.

Data Validation and Correction

The purpose of validation is to produce a survey database suitable for analysis. Such a database must be complete, logically consistent and include documentation of any exceptional conditions present in the data.

The validation program checks the data for each household to insure that:

- all appropriate questions were asked and have valid responses
- no extraneous questions were asked
- responses are logically consistent.

If the program detects any exceptions, it will prepare an exception report showing the data from the questionnaire and error codes identifying the exceptions. This report is then compared to the questionnaire to determine the reason for each exception. The validation error correction manual lists all possible exceptions with a detailed explanation of the cause and suggestions for correction. In general, the correction process consists of checking the responses in the questionnaire and verifying that they were properly recorded and scanned. Often a logically incorrect response may be corrected by examining other data in the questionnaire. When this is not possible, the response is changed to a missing value which indicates that a valid response is not available. Corrections are written on the exception list. This list is used to update the questionnaire database. The error detection and correction process is repeated until only acceptable exceptions, as defined in the validation manual, remain. At this point the cluster folder is stored in the survey archive.

The validation program is run again on the entire database after all individual clusters have been corrected. This final run serves to verify that no serious errors remain to be corrected and provides a summary of the 'acceptable' or non-serious errors that remain in the database.

Tabulation and Other Analysis

Preparation for Tabulation

There are a number of variables needed for tabulation that are not directly available from the questionnaire. These include: household weight factor used to adjust for unequal probabilities of selection of households and to extrapolate the survey results; urban-rural and geographic/administrative region codes used as classifications in the tables; characteristics of the head of household, household size and other individual information aggregated to the household level; child nutritional status variables derived by comparing children's height, weight and age to international standards; and household poverty (welfare) quintile used to classify households by their relative well-being. The definition of the derived variables is done by program when data validation is completed; it is the first step in the tabulation phase of the system.

Tabulation

The CWIQ tabulation plan defines standard tables to be produced within four weeks of the end of fieldwork. These include reference tables, core welfare indicators, and report tables.

Reference tables summarize the responses to all questions by urban-rural and regional classifications. The summaries include the frequency distributions of all responses for discrete variables and the minimum, mean, maximum and standard deviation for continuous variables. Reference tables are a form of quality control. They can be used to: confirm that all variables have valid, reasonable responses; (re)define groupings used in the tables; and verify the contents of the final tables. These tables are not published, but are retained in the statistical office for reference.

The core welfare indicators are summarized in one or two pages. The summary shows the principal indicators and the margin of error (defined as the range of a 95% confidence interval) at the national level and disaggregated indicators for rural, rural poor, urban, urban poor, and regional sub-populations. Simple graphs are used to highlight selected indicators. A sample core welfare indicators summary appears in Annex 8.

The report tables are designed to show the principal results in a relatively compact form; they are the basis for the first survey report. Sampling errors are calculated for all major indicators at the lowest level of disaggregation in the tables. The sampling errors give an indication of the accuracy of the survey estimates and are also published in the main report.

The tables are organized by subject as follows:

1. General results
2. Household situation and characteristics
3. Education
4. Health
5. Nutrition
6. Employment

The tables present the indicators or other characteristics of interest across the columns and the classification or background variables down the rows. A complete list of the report tables appears in Annex 9.

Preparation of analysis files and the survey CD-ROM

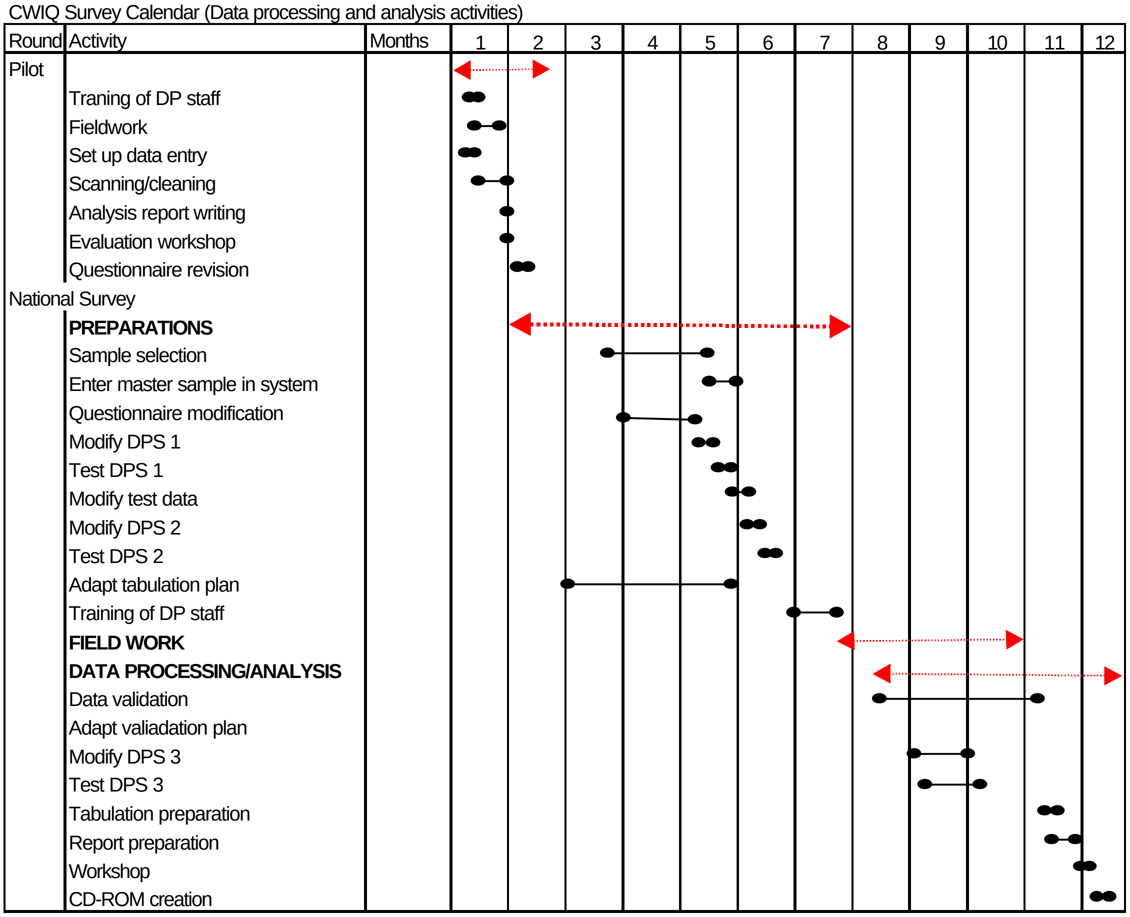
The standard tables and sampling errors are the only outputs produced by the generic data processing system. To allow for further analysis, the survey data are converted to files formatted for processing by general purpose statistical analysis software (SPSS, SAS, etc.). These files and complete survey documentation updated from the generic documentation will be copied to a survey archive on CD. The survey archive will be available for distribution to interested organizations according to data dissemination policies established by the national statistical office.

Data Processing Calendar and Organization

Data processing must be tightly integrated with all other survey activities to produce timely survey results. System development is organized to insure that the system is ready (tested and documented) to process the survey data as soon as it is available. The data preparation, validation and correction programs must be ready when the survey starts so that questionnaires can be processed as they arrive from the field. Processing in parallel with the fieldwork means that a complete, validated survey database will be available as soon as possible after end of fieldwork. The tabulation and other analysis programs must be completely tested and ready for use as soon as the final survey database is ready.

The pilot survey is completely processed to demonstrate the operation of the generic data processing system, to evaluate the questionnaire and to train survey personnel to use the system. After the pilot survey, when the questionnaire is adapted for country requirements, the data processing system is adapted as well. The questionnaire is the first component to be adapted. Questionnaire design requires special software which will not be available in the country until the survey equipment is purchased, so changes to the questionnaire should be determined as early as possible. Data input, validation and correction are done in parallel with the fieldwork, so these components of the system must be adapted, tested and in operation before fieldwork begins. The tabulation programs can be modified and tested during the fieldwork, but must be ready for use as soon as the fieldwork ends. The timetable below shows the sequence of data processing activities and their relationship to the overall survey calendar. There are three stages of system modification identified as DPS1, DPS2 and DPS3. Details of each stage appear below.

Figure 5.1 Schedule of Activities for Data Processing



Adapting the Generic System to Country Requirements

The generic processing system can be used with minimum change to process the pilot survey. A demonstration of the scanning and validation of questionnaires is included in the interviewer training for the pilot survey to demonstrate the new methods and show the special requirements of scanned questionnaires. To further reinforce these lessons, the pilot questionnaires will be scanned and edited daily, and any problems encountered can be immediately communicated to the field staff. At the end of fieldwork, all of the standard tables, except for sampling errors, will be prepared.

In addition to demonstrating the CWIQ methodology, the pilot survey is a test of the CWIQ questionnaire. Although the generic questionnaire was designed to be used with a minimum of change, some modification will be required for each country. In order to realize the benefits of the generic data processing system and achieve the objective of producing rapid results, modifications will have to be limited. In addition to defining poverty predictors, response categories for some of the standard questions will need to be adapted for each country. The pilot survey tables should be consulted to evaluate the suitability of the generic questionnaire categories. Questions can be added where there is room on the page, but the page structure and number of pages (eight) of the generic questionnaire must be respected. If more information is absolutely required, it should be collected in a country specific module(s) for which a separate data processing system will need to be developed. Ideally, all modifications to the questionnaire will be defined by the end of the pilot phase of the survey. This will allow the preparation of a draft questionnaire and a test of the local printing facilities.

The sample characteristics of each survey are country specific. The questionnaire contains only enough information to identify the sample unit (cluster) and household. All other sample information defined at the cluster level (or above) is recorded in a master cluster database. This information must be recorded in the database when the sample is designed. Standard items of information in the master cluster database include:

- Cluster identifier
- Cluster name
- Urban-rural classification
- Geographic/administrative region
- Sample stratum (if sample is stratified)
- Cluster probability of selection or weight
- Number of households listed (if available)
- Number of households selected
- Household weight (“expansion” factor)

The generic data processing system will be adapted in stages to process the modified questionnaire. A test data file for the generic questionnaire has been created to provide a sufficient number of realistic cases to test the validation and tabulation programs. This data must also be adapted for changes in the questionnaire. A model program to adapt the test data is included in the generic system. The modification stages are defined below. The software used for each component is listed in parentheses after the component.

Stage 1 (DPS1) - Questionnaire and scanning system:

- Change questionnaire layout and print new questionnaire (Teleform)
- Change questionnaire data export parameters (TeleForm)
- Prepare test questionnaires
- Test modified scanning system

Stage 2 (DPS2) - Data input, validation and correction (Microsoft Access):

- Modify database definitions
- Write and test program to change generic test data to final questionnaire format
- Update program specifications, modify and test programs
  - Convert scanner output to questionnaire format
  - Transfer data to questionnaire database
  - Validate questionnaires
  - Correct questionnaires
- Update validation manual

Stage 3 (DPS3) - Tabulation (Microsoft Access):

- Modify standard tabulation plan
- Modify definitions of derived variables (including welfare quintiles)
- Update tabulation plan
- Modify and test programs
  - Create derived variables
  - Produce reference tables
  - Produce standard tables
  - Produce sampling errors

The first two stages must be finished before the training course for the national survey. A scanning demonstration during training will accustom the interviewers to the special requirements of scanned questionnaires. A discussion and demonstration of the validation and correction process will show the need for careful interviewing and filling of questionnaires that should improve the quality of the fieldwork. The training course includes several days of practice interviewing and the resulting questionnaires will be processed to reinforce the lessons taught during the course and to evaluate the interviewers.