Data Building and Data Management

This chapter describes the data management procedures used by the SACMEQ study, in which data have been collected from many thousands of pupils sampled from complex target populations in association with different educational systems in Southern Africa. Furthermore, data have been collected at different "levels" (i.e., pupils, teachers, schools) and have been merged, aggregated, disaggregated, etc. prior to the main data analysis. The discussion will therefore focus on the main steps in data management that were required for SACMEQ to take in order to ensure that the quality of collected data is adequate, that data are turned into useful information, and that the more common data management problems are avoided.

1. What is DEM?

In many of the large-scale data surveys, the most common problems in data preparation are: a) coding errors, invalid codes, inconsistent codes, b) lack of automatic saving functions in the data management software, c) lack of smooth switching between data entry, data cleaning, data verification, and data repair modes, and d) problems in transferring from data entry software to data analysis software. The poor data quality introduced by these problems could consequently cause an extensive delay in the further survey implementation.

To address the above problems, SACMEQ used a software system for data management called the DataEntryManager (DEM) programmme. This programmme is available from the IIEP and has been designed specifically to improve data management in educational surveys. This programme has been designed to be used by users with limited experience in computer use. The programme can handle datafiles with more than 1000 variables and data for more than 1 000 000 000 respondents and all datafiles created are fully compatible with the dBASE IVTM standard. It has easy-to-learn features for data entry, editing, validation, and data verification and comes with integrated file management and reporting capabilities. Using this programme, deviations of data values from pre-specified validation criteria or data verification rules can be detected quickly, thereby allowing the user to correct errors quickly after the original survey materials arrive at the survey office. The manual for the DEM programme (see *DEM User's Guide*) describes how to create new

datafiles or to modify the structure of existing datafiles, and how to change coding schemes and range validation criteria for variables. The manual also contains a tutorial in which the user can learn interactively how to transform a questionnaire into an electronic codebook, how to set up a datafile, how to enter data into this datafile, and how to backup your data on diskettes.

2. Data Sources and Datafiles

From the dummy tables prepared based on the "high" priority policy questions (see Chapter 2), the SACMEQ NRCs decided to collect information from three different levels of information source: (a) school level, (b) teacher level, and (c) pupil level.

At the school level, datafiles for School Head Questionnaire (see Chapter 5) and the School Form have been constructed. The datafile for the school form included the following items: (i) the official identification number of the school; (ii) the name, full address, and telephone number of the school; (iii) the name of the person co-ordinating the assessment in the school; (iv) the number of classes in the target population in the school; (v) the number of pupils in the target population in the school; (vi) the number of teachers in the target population in the school; and (vii) name and identification of the teachers in the target population in the school.

At the teacher level, a datafile for Teacher Questionnaire (see Chapter 5) has been constructed. At the pupil level, two datafiles have been constructed: one for the Pupil Booklet (including Pupil Reading Test and Pupil Questionnaire) and one for the Pupil Name Form. The datafile for the Pupil Name Form provided information on: (i) the identification number and the name of pupils; (ii) the date of birth and sex of pupils; (iii) the class of pupils; (iv) the participation status of pupils in the test and questionnaire administration; and (v) the pupils who have been excluded from the testing.

3. Construction of Structure Files

Total of five structure files have been constructed; they are for Pupil Booklet, Teacher Questionnaire, School Head Questionnaire, School Form, and Pupil Name Form. Based on the questions in these instruments, necessary variable(s) were defined with detailed coding specifications.

For each variable, the following essential pieces of information were defined: (i) unique variable name; (ii) variable type (a fixed set of categories or open-ended numerical codes); (iii) variable length (number of spaces taken for this variable); (iv) number of decimal places; (v) where the information on this variable can be found in he instrument; (vi) variable label which describes the variable; (vii) code for "missing" data' (viii) code for "not administered" data; (ix) "default" code for newly created record; (x) valid range; (xi) "modification" indicator; (xii) "carry" indicator; and (xiii) variable class.

In order to code for the missing data above, it should be noted that some distinctions between different instances of missing data were made before the data were entered into the datafile. Later on, there were other distinctions which were derived when the data were being processed.

The "missing/omit" codes referred to questions/items which a respondent should have answered but which he/she either did not answer or which were answered in an invalid way (though sometimes a finer distinction between these categories may be required). Some obvious reasons for assigning this code were: (a) no response; (b) two or more responses; and (c) response unreadable.

The "not administered" codes were assigned when data were not collected for an observation on a specific variable. There were some obvious cases when this code was used: (a) respondent not present; (b) booklet not received; (c) item left out or misprinted; and (d) item mistranslated.

4. Valid Code Ranges

For each non-categorical (open-ended numerical) variable, a valid range has been established. This was done in order to control the quality of incoming data. Valid ranges for the variables used in Pupil Questionnaire, Teacher Questionnaire, and School Head Questionnaire have been presented in Appendix E.

5. Codebooks

A set of detailed instructions describing how data must be coded and how results are stored in computer readable form is usually referred to as the codebook. For each datafile created with the DataEntryManager programme, the programme maintained an electronic codebook which contains all technical information required to define the file structure, the coding scheme, the data verification rule, and quality standards for the datafile. Whenever variables were modified, the programme updated the electronic codebook automatically.

In Appendix F, the codebooks for the Pupil Booklet, Teacher Questionnaire, and School Head Questionnaire have been presented. The different pieces of information that are contained in these codebooks are described below:

(a) the first column in the codebook (Var. No.) presents a sequential number for each variable in the Codebook;

(b) the second column (Quest. No.) presents an identification of the background question and its location in the instruments;

(c) the third column (Variable Name) presents the variable name;

(d) the fourth column (Variable Label) presents the variable label;

(e) the fifth column (Code R:Recode) presents the codes for the responses, and the recodes for variables for which recoding is necessary and where recoding is not covered by the general notes on recoding. Whenever actual numerical data are supplied in the response

to the questions, this is indicated by the keyword "VALUE". The missing-code presented in the codebook indicates "missing/non-response" values. The "not administered" code presented in the codebook indicates "not administered" values;

(f) the sixth column (Option) presents the response phrase (or an abbreviation of it) that corresponds to the code. For variables that contain actual numeric data, it contains an explanation and the permitted range of the value to be entered;

(g) the seventh column (Location/Format) presents the location and format of the variable in the raw datafile. A variables format is the pattern used to write each value of the variable. It consists of the variable type, the first column in the raw datafile that is assigned to the variable, the last column the variable occupies in the raw datafile, and the length and the number of decimal places. In the seventh column of the codebook the first two numbers refer to the position of the first and last digit of the value of a variable within a record. "C" and "N" indicate the variable type (where N refers to open-ended numeric variables and C refers to categorical alpha-numeric values). The third number refers to the length (where the numeric code refers to the length of the value and the number of decimal places associated with the values) of each variable.

(h) for some background variables, a comment indicates special coding instructions.

6. Data Entry and Data Cleaning

Each National Research Co-ordinator has received the above structure files and codebooks prior to the data collection. He/she has then installed on several computers at their Ministry and trained data enterers. After the data were returned from the schools, the data were recorded in computer readable form.

The programme only allowed the data enterers to enter those variables which match the criteria which have been specified in the codebook. For example, if the code "3" is entered for the variables for pupil sex, the programme would reject this value. This is because only the codes "1" for "boy", "2" for "girl", "9" for "missing", and "8" for "not administered" have been defined.

If, for an open-ended variable (variable type "N") a data value is entered which is outside the range specified in the codebook, then the programme would alert the data enterers.

A critical step in the management of survey data is the verification of the data. It must be ensured that the data are consistent and conform to the definitions in the codebook and are ready for analytical use. This step is often referred to as data cleaning. Data cleaning steps taken for SACMEQ were as follows:

(a) the verification of returned instruments for completeness and correctness;

(b) the verification of identification codes against field monitoring and survey tracking instruments;

(c) a check for the file integrity, in particular, the verification of the structure of the datafiles against the specifications in the codebook;

(d) the verification of the identification codes for internal consistency;

(e) the verification of the data variables for each pupil or teacher against the validation criteria specified in the codebook;

(f) the verification of the data variables for each pupil and teacher against certain control variables in the datafiles;

(g) the verification of the data obtained for each respondent for internal consistency (for example, the responses for questions which were coded through split variables the answers to these can be cross-validated);

(h) the cross-validation of related data between respondents;

(i) the verification of linkages between related datafiles, especially in the case of hierarchically structured data; and

(j) the verification of the handling of missing data.

7. Database Construction and Database Management

In order to use the collected information more efficiently, the cleaned datafiles were integrated into a SACMEQ database system. This database system is a structured aggregation of data-elements which can be related and linked to each other through specified relationships. The data-elements can then be accessed through specified criteria and through a set of transaction operations, which are usually implemented through a data-retrieval language. In such a database system the links between the physical data stored in the computer, their conceptual representation, and the views of the users on the data are implemented through a database management system.

The construction of the database system included the following steps:

(a) create a score file which contain the test scores for all the sub-dimensions for the test;

(b) create a variable which carry the proper "weight" for each pupil in order for the collected sample to be proportionate to the number of pupils in each stratum;

(c) merge the pupil datafile with score file, teacher datafile, and school head datafile; and

(d) construct derived variable using the existing variables.