END-DECADE

MULTIPLE-INDICATOR DATA PROCESSING MANUAL

IMPS

Contents

Overview

Data processing personnel

The data processing team for a MICS survey includes four types of personnel: questionnaire administrators, data entry operators, secondary editors and a data processing supervisor. Each position has distinct responsibilities and combining them is likely to damage the quality of your data.

Questionnaire administrators

The questionnaire administrators check the clusters that arrive from the field. They check that all of the questionnaires are present and ready to be entered. If there are missing questionnaires, they must contact the fieldwork team and try to find them. If the missing questionnaires cannot be found, the questionnaire administrators must resolve the problem.

In the case of household questionnaires, this means requiring the fieldwork team to redo the household interview. If this is impossible, the questionnaire administrators must add a blank household questionnaire with the result code 5.

In the case of women's and children's questionnaires, this means requiring the fieldwork team to redo the missing questionnaires. If this is impossible, the questionnaire administrators must change the completed questionnaire totals on the household cover sheet.

Data entry operators

The data entry operators enter the data. They should have prior data entry experience and be familiar with the questionnaires. Before beginning data entry in earnest, a two or three day training must be held. By the end of the training, the data entry operators should be comfortable with the data entry program and aware of their responsibilities.

Secondary editors

Secondary editors investigate complex inconsistencies discovered during the data entry process. They must have an excellent grasp of the questionnaires and the goals of the survey. Editing guidelines are provided to aid them during the secondary editing process.

Data processing supervisor

The data entry supervisor oversees the data processing system. The data entry supervisor must have an excellent grasp of the questionnaire and programming skills in the data entry package and SPSS. The data entry supervisor also resolves problems encountered by the data entry operators.

Data processing equipment

The list below shows the equipment that is necessary for data processing.

- Data entry machines
- A supervisor's machine
- A secondary storage device (e.g., a Zip or Jaz drive)
- Diskettes
- A printer

- Paper
- Toner cartridges/printer ribbons
- Surge protectors
- Uninterrupted power supplies
- Green pens
- A data entry room
- An editing room

The data entry machines should have at least 486 processors, Windows 95 or higher, 16 megabytes of RAM or more, 500 megabytes of hard disk space or more and a 3.5'' floppy diskette drive. The supervisor's machine should have a 300 Mhz processor, Windows 95 or higher, 64 megabytes of RAM or more, 500 megabytes of hard disk space or more, a 3.5'' floppy diskette drive and a CD-ROM drive. The secondary storage device should be connected to the supervisor's machine.

The Uninterrupted power supplies are essential if the country in which you are working suffers from power outages. The green pens are for the data entry clerks and the office and secondary editors. They should be used whenever a questionnaire is modified. The green ink distinguishes their changes from the original value recorded by the interviewer and the changes made by the fieldwork team.

The data entry room should be large and cool. Each data entry agent should have space for their machine and the questionnaire they are working on. The editing room is for the office and secondary editors. It must contain shelves or cupboards in which the questionnaires can be stored in an organized fashion.

Data processing design

MICS data processing is organized around clusters. The data for each cluster are entered into a set of data files whose names include the cluster's number. This approach breaks data processing up into discrete segments and keeps the size of data files to a minimum.

A MICS survey has two data processing phases: primary and secondary. The goal of primary data processing is to produce clean, edited data files. Primary data processing comprises the following steps.

- Data entry
- Structure check
- Verification
- Secondary editing
- Backup of data

The goal of secondary data processing is to produce data files to be used for analysis, including tabulations. Secondary data processing comprises the following steps.

- Calculation of weights
- Recoding of variables
- Tabulation

The flow chart on the following page illustrates the flow of primary data processing.

```
┌─────────────────┐
│  Main Data Entry │
└─────────────────┘
         │
         ▼
┌─────────────────┐
│  Structure Check │ ◄──────────────────────────────────┐
└─────────────────┘                                     │
         │                                              │
         ▼                                              │
    ╱‾‾‾‾‾‾‾╲                                    ┌──────────────────┐
   ( Structure  )        No                      │ Correct Data File │
    ╲ okay? ╱ ──────────────────────────────►    └──────────────────┘
    ╲‾‾‾‾‾╱
         │
         ▼
┌─────────────────────┐
│ Verification Data Entry │
└─────────────────────┘
         │
         ▼
┌─────────────────────┐
│ Compare Main and    │ ◄──────────────────────────────┐
│ Verification Data Files │                              │
└─────────────────────┘                                 │
         │                                              │
         ▼                                              │
    ╱‾‾‾‾‾‾‾╲                                    ┌──────────────────┐
   ( Differences )        Yes                    │ Correct Data Files │
   (  found?   ) ──────────────────────────►     └──────────────────┘
    ╲‾‾‾‾‾╱
         │
         ▼
┌─────────────────────┐
│ Backup Raw Data File │
└─────────────────────┘
         │
         ▼
┌─────────────────┐
│ Secondary Editing │ ◄──────────────────────────────────┐
└─────────────────┘                                     │
         │                                              │
         ▼                                              │
    ╱‾‾‾‾‾‾‾╲                                    ┌──────────────────┐
   ( Inconsistencies )    Yes                    │ Correct Data File │
   (    ?    ) ──────────────────────────────►   └──────────────────┘
    ╲‾‾‾‾‾╱
         │
         ▼
┌─────────────────────┐
│ Backup Final Data File │
└─────────────────────┘
```
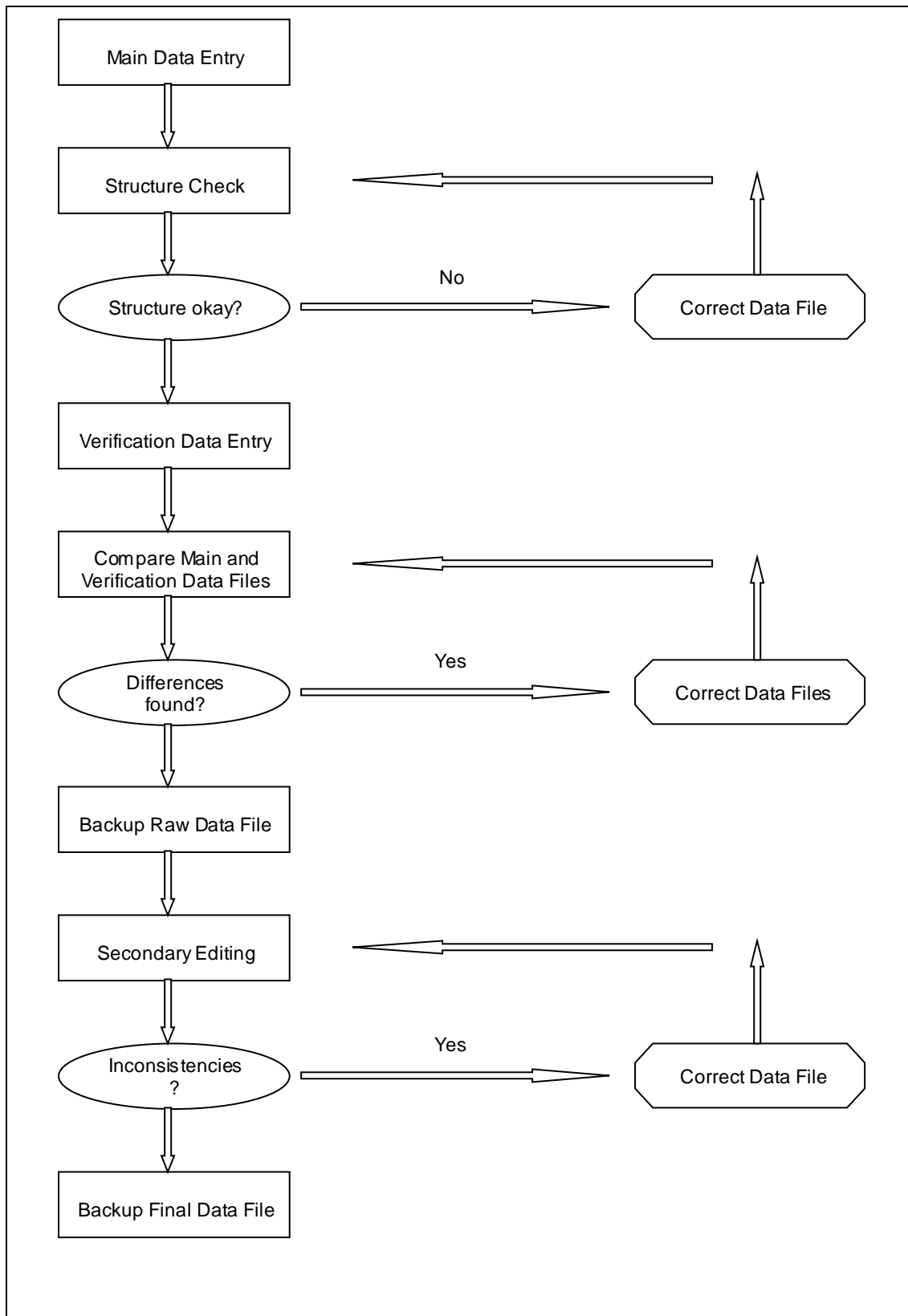
Main data entry

The data for each cluster are entered twice. The first time the data is entered is called main data entry and the second time is called verification data entry. One data entry agent must do the main data entry and another data entry agent must do the verification data entry. Under no circumstances should the same data entry agent enter both the main and verification data.

The data entry program performs basic consistency checks. These checks are designed to catch egregious questionnaire errors and data entry errors. Complex inconsistencies are not addressed until secondary editing.

Structure check

It is essential that all questionnaires be entered and that their identification information be correct. The structure check ensures that both of these requirements are met. Only the structure of the main data entry needs to be checked because any structural errors in the verification data will be discovered during the verification process.

Verification

When main and verification data entry are finished, the two files are compared and a list of differences is produced. The two data entry clerks resolve differences by checking the questionnaires for the correct values. They then correct the values in their data files and the verification program is rerun. The process is continued until the data files are identical. When the verification process is finished, the data should contain exactly what is written on the questionnaires.

Raw data backup

When the data have been verified a backup copy is made before beginning secondary editing. In addition to guarding against accidental data erasure, this backup allows survey personnel to review the changes made during secondary editing. It is recommend that the raw data be backed up to a secondary storage medium (e.g., a zip disk). If this not possible, make a backup copy of the raw data in a separate directory on the hard drive of the supervisor's machine.
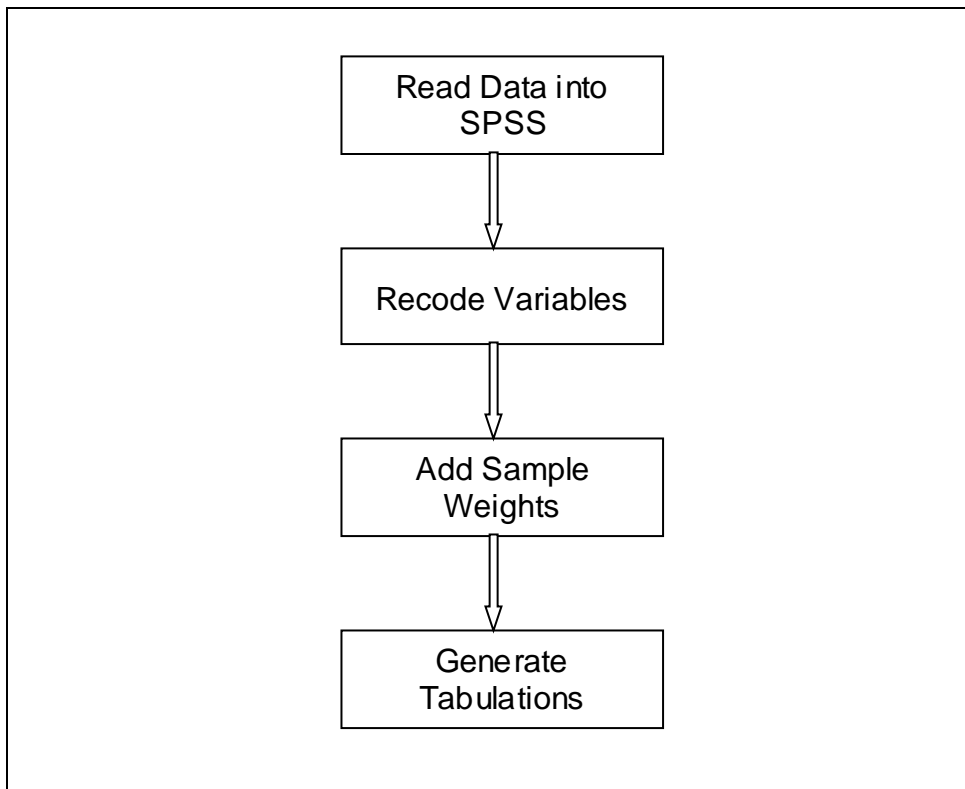
Secondary editing

The goal of secondary editing is to investigate complex inconsistencies. The goal is not to systematically correct these inconsistencies. Inconsistencies should only by resolved if they concern essential variables (e.g., ages), if there is clear evidence of interviewer error or if the inconsistency is due to data entry error.

Secondary editing can involve several iterations because resolving one inconsistency often introduces others.

Final data backup

When secondary editing is complete, the final data are backed up. It is recommend that the final data be backed up to a secondary storage medium (e.g., a zip disk). If this is not possible, make a backup copy of the final data in a separate directory on the hard drive of the supervisor's machine.

The chart below illustrates the flow of secondary data processing.

```
┌─────────────────┐
│  Read Data into │
│      SPSS       │
└─────────────────┘
         │
         ▼
┌─────────────────┐
│ Recode Variables│
└─────────────────┘
         │
         ▼
┌─────────────────┐
│   Add Sample    │
│     Weights     │
└─────────────────┘
         │
         ▼
┌─────────────────┐
│    Generate     │
│   Tabulations   │
└─────────────────┘
```

Concatenating data

The first step in secondary data processing is concatenating all of the cluster files into one national file.

Exporting data

The second step in secondary data processing is exporting the data from the data entry package into SPSS. This includes transferring any variable or value labels.

Recoding variables

Once the data have been exported to SPSS, the process of creating analysis files begins. The first step is to create new variables and to recode existing variables.

Add sample weights

Sample weights are calculated using an Excel spreadsheet. When all the required analysis variables have been created, the sample weights are added to the data files.

Generate tabulations

When analysis files have been created, SPSS programs generate the standard MICS tabulations.

Control Sheet

The control sheet is used to manage the flow of primary data processing. It is stored in an Excel file in the SUPER directory named CONTROL.XLS. The first column contains all of the cluster numbers. Columns B, D, E, G, H, I and J contain space for writing the dates on which the primary data processing tasks are completed. Column C contains space for writing the number of the data entry operator who was assigned main data entry. Column F contains space for writing the number of the data entry operator who was assigned verification data entry.

Before the start of data entry, print a copy of CONTROL.XLS. This will serves as a paper backup of the control sheet. Each time a task is completed, update both the paper copy and the file. The second to last row of the control sheet contains the number of clusters for which each task has been completed. The last row of the control sheet contains the percent of clusters for which each task has been completed. Use these two rows to track the progress of the data processing system.

CONTROL.XLS needs to be modified for your survey. Modify the cluster numbers so that they match the cluster numbers for your survey, adding or deleting rows as necessary. Make sure that cells are correctly formatted (e.g., cells that are to contain dates are defined as date cells). Finally, check the formulas in the last two rows and update them if necessary.

Cluster summary sheet

A cluster summary sheet should be maintained for each cluster of questionnaires. It should be a large, heavy piece of paper with the cluster number written on it in large, bold numerals. The cluster summary sheet should be folded over the end of the package of questionnaires so that the cluster number is displayed clearly. The following fields should be printed on the top of the sheet.

Cluster Number   _____
Total Household Questionnaires   _____
Total Households Completed   _____
Total Women Questionnaires   _____
Total Children Questionnaires   _____

The questionnaire administrator should complete the fields when he or she receives the cluster.

Data entry directory structure

The diagram below shows the directory structure on a data entry machine.

```
MICS
        EPIINFO/IMPS/ISSA
                DATA
                ENTRY
                VERIFY
```

EpiInfo/Imps/Issa

This directory, which is named for the data entry package being used, contains the entry menu.

Data

The data directory contains the results of main data entry.

Entry

The entry directory contains the data entry programs.

Verify

The verify directory contains the results of verification data entry.

Supervisor directory structure

The diagram below shows the directory structure on the supervisor's machine.

```
MICS
        EPIINFO/IMPS/ISSA
                BACKUP
                EXPORT
                FINAL
                RAW
                SPSS
                SUPER
                WEIGHTS
```

EpiInfo/Imps/Issa

This directory, which is named for the data entry package being used, contains the supervisor menu.

Backup

The backup directory contains a backup of data files that have been structurally checked and verified but not edited.

Export

The export directory contains the programs used to export the data from the data entry package to SPSS.

Final

The final directory contains a backup of data files that have been structurally checked, verified and edited.

Raw

The raw directory contains the data files that have been transferred from the data entry machines.

Spss

The spss directory contains the programs that create analysis files and the tabulation programs.

Super

The super directory contains the programs that do structural checks, verification and secondary editing.

Weights

The weights directory contains the spreadsheet that calculates sample weights.

Data description

There are three types of MICS questionnaires: household, women and children under five. The women's and children's questionnaires each correspond to a single unit of analysis: a woman and a child respectively. The household questionnaire contains two units of analysis: the household and the household's member.

Each of the questionnaire types is stored in a separate data file and each file uses a separate data description (DD) file. The data description files are named as follows.

Questionnaire	Data Description
Household	MICSHH
Women	MICSWM
Children	MICSCH

IMPS therefore has three DD files for MICS. These DD files should only be modified using the IMPS user interface.

Modules

The MICS questionnaires are made up of modules. Countries can add or remove modules, depending on their data needs. Countries are discouraged from significantly altering the contents of a module. The modules available for each questionnaire are listed below. The codes listed below are also used as record identifiers in the data.

Questionnaire	Module	Code
Household	Household information	HI
	Household listing	HL
	Education	ED
	Child labor	CL
	Water and sanitation	WS
	Salt iodization	SI
Women	Women's Information panel	WI
	Child mortality	CM
	Tetanus toxoid	TT
	Maternal and newborn health	MN
	Contraceptive use	CU
	HIV/AIDS	HA
Children	Birth registration and early learning	BR
	Vitamin A	VA
	Breastfeeding	BF
	Care of illness	CI
	Malaria	ML
	Immunization	IM
	Anthropometry	AN

Each module in IMPS is represented by a record type in the data definition file. The names of these records all begin with "MOD", followed by the module code. For example, the record for the education module in the household questionnaire is "MODED."

Variable naming conventions

Variables are named for the module that they are in and the number of the question whose response they contain. For example, question 8 in the household listing is named HL8.

Some questions are split into two or more parts, with each part being prefaced by a letter. Each part of the question is a separate variable. The names of these separate variables include the letters that distinguish the parts of the question. For example, question 3 of the women's information panel has two parts, 3A and 3B. Question 3B is named WI3B.

Some questions have two or more parts to the response categories. When these questions concern data, the letters D (for day), M (for month) and Y (for year) are appended to the base variable to create separate variables. In question 3A of the women's information panel, for example, both a month and a year are required. These two variables are named WI3AM and WI3AY.

There are two questions in the questionnaires in which the first part of the response is the form of the response and the second part is the response: question 6 in the maternal and newborn health module of the women's questionnaire and question 3 of the anthropometry module of the children's questionnaire. In both cases, the variable giving the form of the response has the letter A appended to its name while the variable containing the response has no letter appended (e.g., MN6A and MN6, AN2A and AN2).

There are a number of questions in the questionnaire that allow for multiple responses. For these questions, each response has its own variable. The variables' names are made up of the module code, the question number and the letters A through K, which correspond to the response codes 1 to 11. An exception is the response "don't know" which uses the letter Z in all cases. For example, question 4 of the malaria module in the children's questionnaire has variables named ML4A, ML4B, ML4C, ML4D, ML4Z.

Coding conventions

The standard programs use standard coding for certain responses. The response "don't know" is always coded as a nine with leading nines. Questions that are not applicable to a respondent are always coded as an eight with leading nines. Questions with a missing response (i.e., the interviewer did not record a response to an applicable question) are always coded as a seven with leading nines. Responses that are inconsistent with other data in the questionnaire are always coded as a six with leading nines. The table below summarizes the standard coding conventions.

Response	Variable length			
	One character	Two characters	Three characters	Four characters
Don't know	9	99	999	9999
Not applicable	8	98	998	9998
Missing	7	97	997	9997
Inconsistent	6	96	996	9996

All the response categories for a variable should contain the same number of digits. If a question requires two digits, codes with values less than 10 must have leading zeros. For example, if a question uses codes 1 through 12, the codes on the questionnaire must be 01-12 and not 1-12.

Because the codes 7, 8 and 9 are reserved for special use, any question that requires more than 6 response categories must have 2-digit responses categories with leading zeros (e.g., 01, 02, 03, 04, 05, 06, 07, 97 and 99).

Ranges

Most of the variables in the MICS questionnaires have defined ranges. Only values within the specified ranges can be entered. This reduces the probability of data entry errors. The only variables without defined ranges are variables for which it is impossible to create restrictive ranges. The limits of household number, for example, vary so much from cluster to cluster that any defined ranges would catch only a small number of errors.

Identification variables

Every questionnaire must have a series of variables that uniquely identifies it. The variables that identify a questionnaire are known as the identification variables. The table below lists the questionnaire types and their identification variables.

Questionnaire type	Cluster number	Household number	Line number
Household	HI1	HI2	
Household listing	HI1	HI2	HL1
Women	WICLNO	WIHHNO	WILNNO
Children	CHCLNO	CHHHNO	CHLNNO

MICSHH.DD

This data definition file contains the definitions of all modules in the household questionnaire. The data for each household is uniquely identified by its cluster number (HI1) and household number (HI2), each of which is three characters long. These two fields therefore occupy columns 1 through 6 in the data.

MICSWM.DD

This data definition file contains the definitions of all modules in the women's questionnaire. The data for each women is uniquely identified by its cluster number (WICLNO), household number (WIHHNO), and the woman's line number in the household listing (WILNNO). As in the household data, the cluster and household numbers are each 3 characters long. The woman's line number is 2 characters long. The entire identification therefore occupies columns 1 through 8 in the data.

MICSCH.DD

The data definition file for the children's questionnaire contains the definitions of all modules in the children's questionnaire. Similar to the women's data, data for each child is uniquely identified by its cluster number (CHCLNO), household number (CHHHNO), and the child's line number in the household listing (CHLNNO). As in the women's data, the cluster and household numbers are each 3 characters long, and the child's line number is 2 characters long. The entire identification therefore occupies columns 1 through 8 in the data.

CONCOR data entry programs

IMPS data entry applications (AP files) contain data entry forms and commands for skip patterns as required by the questionnaire. The skip patterns control the data entry path (the path that a data entry operator follows through the questionnaire).

The following options should be selected for the data entry applications:

```
            Application Options

  Confirm data entry      NO
  Force decimals          YES
  ID checking             BATCH ONLY
  Automatic skips         YES
  Bypass sequence items   YES
  Beep type               00
  Allow force key         NO
  Keycode application     NO
  Sample verification     Start: 001
                           Freq: 001
```

Skips

The MICS questionnaires make abundant use of skips. Skips are instructions on the questionnaire that tell the interviewer to skip all the questions between the current question and a question later on in the questionnaire. Skips on a questionnaire must be matched by skips in the corresponding AP CENTRY application file. Skips in a data entry program define the data entry path.

IMPS does not strictly enforce the data entry path. Data entry operators can enter and change fields that are not on the data entry path by backing up into a previously blank field. While this cannot be avoided, it is strongly advised to leave the "Allow force key" option off to avoid the entry of values that are out of range as defined in the data definition.

Many of the skip patterns in the data path are checked in the corresponding CN CONCOR file. Some of the details of these files are discussed below.

Error messages

Error messages are displayed by the CN application whenever a skip pattern has been violated, or the structure of the data is incorrect. The error messages are numbered and a master list of the error message number exists. If you add error messages, make sure that you do not reuse a number, and that you document the meaning and recommended response to that message.

Multiple response questions

There are a number of questions in the questionnaire that allow for multiple responses. For these questions, each response has its own variable. Each variable has three possible responses – 0 for no, the responses code on the questionnaire, and 7 for missing data.

For example, the variable MN2 records whom a woman saw for antenatal care before her last birth. The response code 2 is associated with the response nurse/midwife. The variable for this response is named MN2B. The allowed entries are 0, 2 or 7.

For all of these multiple response questions, the use of missing values must be consistent. Either all of the individual variables related to any given question must be missing, or none of them can be missing.

Structure and Consistency Issues

Structural issues are the types of problems in data entry that are the most important to correct. These issues relate to the numbers of people who are eligible for women's or children's questionnaires, compared to the actual numbers of these questionnaires, and the presence or absence of modules based on responses in the household questionnaire.

These are checked after each questionnaire, for example in MICSHH.CN, and then more thoroughly by the supervisor, in CHECKHH.CN. Inconsistencies are also checked in MICSHH.CN, and many others are also checked in EDITHH.CN. Corresponding WM and CH files are also used, and will be discussed later.

MICSHH.CN

Understanding how the CN files work is essential to understanding the MICS data entry programs. Most of this application is dedicated to enforcement of structural rules. It first checks that the cluster number is among those allowed by the sample design. This should be modified according to the specific design of your survey. It then checks to be sure the households are entered in ascending order.

This application then counts the number of household members who are eligible for the education and child labor modules, and compares that total to the number of modules that were actually entered. Any problems discovered that are related to this or to any other structural issues should be resolved, even if this means re-entering a substantial number of questionnaires.

MICSWM.CN

There are many structural checks in this application, as mentioned above. There are also a large number of consistency checks. As in the household application, the cluster number is checked to be sure it is an acceptable value based on the sample design. Again, this should be modified to suit your survey.

The application then checks for the presence of modules, including the women's information, child mortality, tetanus toxoid, maternal and newborn health, contraceptive use, HIV/AIDS, and newborn health modules. Structural checks are also done within many of the procedures for the individual modules.

These modules also include consistency checks, most of which should not be changed. For example, the date of birth must be feasible based on the month and day entered, so April 31 should not be allowed. The application also accounts for the year 2000 as a leap year in the permissible days in February.

The information stored in the birth history is used in many places of the questionnaire for further data entry, and is pivotal to many of the analyses for which your data will be used. It is extremely important that the data in this section be as accurate and consistent as possible without falsifying the data or making up any new data. The numbers of boys and girls at home, living elsewhere, and who have died, as first given by the respondent, must match the numbers of such children in the birth history.

Dates of birth as given in the birth history must also be as complete and accurate as possible, especially the dates of first and last births. Many checks are performed on these dates to ensure consistency. You should not alter these checks.

Several questions in the women's questionnaire are in fact multiple response questions which have been split into separate questions for each possible response. These include antenatal care, delivery assistance, and current method of family planning. All of these require special treatment for missing responses: the use of missing is valid only if *all* of the responses for that question are missing. For this reason, the application checks that the total number of responses of "missing" is either zero or equal to the number of questions. Do not make any changes to the syntax of these checks.

MICSCH.CN

This application starts similarly to the household and women's applications: it checks that the cluster number is valid, and that the households have been entered in order. It then checks that the numbers of occurrences of the birth registration, vitamin A, breastfeeding, care of illness, malaria, immunization, and anthropometry modules are included exactly one time for each case. If your survey excludes any of these modules, remove that module from this section of the application. The application then checks that the child's line number is different from the line number of the child's caretaker.

The application then checks for consistencies in multiple response questions, as in the women's module. There include source of care for illness, symptoms that indicate child should be taken to a health facility, medicine taken at health facility, and medicine given for malaria or fever during illness. Each of these is checked to be sure that there are either no missing values, or all of the answers are missing.

Structure checking

Structure checking is an essential part of the MICS data processing approach. It is essential that the data be structurally sound. The data entry programs do not enforce structural coherence (there are several reason for this, but the most important is that it would constrain data entry too much) so a structural checking program must do so.

Assumptions

Before examining the structure checking program, consider the lists below that layout some of the structural assumptions that are being checked. The lists are not exhaustive, but they do illustrate the general principles on which structural checks are based.

Household

- The number of households should equal the sum of the number of incomplete and complete households.
- The number of records in the household data file should equal the number of households implied by the household listing data file.

Household listing

- The number of household members listed on the household cover sheet should equal the number of household member records in the household listing data file.
- The number of household members who are women age 15 to 49 should match the number of eligible women on the household cover sheet.
- The number of household members who are children under the age of 5 should match the number of eligible children on the household cover sheet.

Women

- The number of women's records should match the number of interviewed women on the household cover sheet.
- There should be a one-to-one correspondence between records in the women's data files and records in the household listing data file for women age 15 to 49.

Children

- The number of children's records should match the number of completed children interviews on the household cover sheet.
- There should be a one-to-one correspondence between records in the children's data files and records in the household listing data file for children under the age of 5.

Notice that many of the assumptions above are true for both the household and the cluster. For example, the number of eligible women on the household cover sheet should equal the number of female household members age 15 to 49. At the same time, the sum of eligible women for all households in the cluster should equal the number of female household members age 15 to 49 for the entire cluster.

Structure Checking Programs

Structure checking is done using CHECKHH.CN, CHECKWM.CN, and CHECKCH.CN. Together, these three applications check the structural coherence of the household (include the household listing), women's and children's data files. On completion, these applications generate detailed reports with the numbers of questionnaires that have been entered and their result codes.

The workings of the three parts will not be explained here in detail. We will focus on what the programs are doing instead of how they do it. The programs are complex and do not lend themselves to simple explanations. The best way to understand them is to study them carefully when you understand what they are doing and why.

CHECKHH.CN

This application examines the household data, looking for numbers of questionnaires, numbers of household members who are eligible for other questionnaires, and comparing the numbers entered on the cover sheet to the sums in the household listing.

First, the application counts the number of households for whole cluster, with subtotals for the number of complete and incomplete questionnaires. It then checks that there is exactly one coversheet for each household. For completed household questionnaires, it checks that the total and eligible numbers of household members given on the cover sheets are the same as those computed within the household listing. It also checks that the numbers of modules in the data for each household is correct.

Within the household listing, this application compares each member's eligibility with the information on that member, to determine if the eligibility is correct.

The results of these comparisons are displayed in a report that is described in detail below.

CHECKWM.CN

The primary objective of this application is to ensure that the numbers of modules in each case are correct. It also checks that the women's eligibility in the household listing is correct.

CHECKCH.CN

This application serves the same purposes as CHECKWM.CN, but for the children's questionnaire. In addition, it checks the consistency of the caretaker's line number.

Structure Checking Reports

Part one

The first part of this report displays counts of the number of questionnaires expected and compares the results to the number of questionnaires in the data files. The output below shows an example based on one household.

```
                              MICS Data Structure Check
        Cluster:   3

            Households      |Household Members|     Women        |     Children
        Total  Comp  Incomp|  Cover     Found | Elig. Int. Found | Elig. Int. Found
          1     1      0   |    6         5   |   2    2    2    |   0    0    1

        ** Total household members incorrect
        ** Total children interviewed incorrect
```

There are two structural problems in this example. First, the household cover sheet records 6 household members but only 5 household listing records exist in the data file. Second, there is a case in the child's data file, but the household cover sheet records no children under the age of 5 and no children's interviews.

These errors can be seen from the summary table and from the two error messages.

Part two

The second part of report provides a summary of households and their members. Information about the household is displayed first followed by information about the household's members, if any. Information is displayed about all households in the data files, even if they are structurally correct. The output below continues the one household example used in part one.

```
        Household      HI10 |  HIMEM        |HI11    HI12     |HI13    HI14
        Household member    |          HL1  | HL5      WILNNO| HL7       CHLNNO

            3001   res=1 |    6          | 2      2        | 0      0
            300101       |          1    | 0        .      | 0          .
            300102       |          2    | 2        2      | 0          .
            300103       |          3    | 3        .      | 0          .
            300104       |          4    | 0        .      | 0          .
            300105       |          5    | 0        .      | 0          5
        ** 0210 Interviewed child not eligible
        ** Number of household members incorrect
        ** 0209 Number of interviewed women incorrect
        ** 0208 Number of interviewed children incorrect
```

HI10 is the result of the household interview. In the example, the household interview was completed (res=1). HIMEM is the number of household members. Notice that it is followed by only 5 household members; the household member with line number 6 is missing.

HI11 is the number of eligible women. HL5 is a household variable that reflects eligibility for the women's interview. Two household members are recorded as eligible, a total that matches HL5. HI12 is the number of interviewed women. WILNNO is the line number for a record in the women's data file. Only one household member has a value meaning that one women's questionnaire is missing for this household, one women's questionnaire for this cluster has an incorrect line number or the value of HI12 is incorrect.

HI13 is the number of eligible children; HL7 is the line number of an eligible child's caretaker. HI14 is the number of complete children's interviews. CHLNNO is the line number for a record in the children's data file. One such record exists despite the value of HI14 (0) meaning that the eligibility code of one of the household members is incorrect and the values HI13 and HI14 are incorrect or the household identification for the child's questionnaire is incorrect.

The four messages that follow the household summary summarize the errors mentioned above. Messages one and four derive from the same source.

Part three

Part three is similar to part one. The difference is that part one's counts are based on what is present in the data files while part three's counts are constrained by some basic consistency checks. The output below continues the one household example used in part one.

```
         Households      |Household Members|      Women        |      Children
    Total  Comp  Incomp|  Cover    Listed | Elig.  Int.  Found | Elig.  Int.  Found
      1     1      0   |    6        5    |   2     2     1    |   0     0     1

    ** Total household members incorrect
    ** Total women interviewed incorrect
    ** Total children interviewed incorrect

    ** Total women's questionnaires inconsistent
```

The difference between these two approaches can be seen in the found column of the women's section of the table. Part one found 2 women; part three found 1 woman. Part one counted the number of women's records in the women's data files. Part three counted the number of household listing records who had a matching record in the women's data file.

The missing women's questionnaire had an incorrect line number: 9 instead of 3. Because no household listing record has a line number of 9, this women's record could not be matched to a household listing record and was not counted. The error messages that follow the table summarize the errors that can be observed in the table.

Note that the totals that are used in part three are computed in part two; only the output is created in part three.

Verification

Verification is done using the IMPS verify module. The main data entry file (which has been copied from a diskette) is compared to verification data as it is entered. Verification is done during this data entry, and differences are reported as they are found. The data entry operator must then select the correct response before continuing.

After verification of an entire cluster is completed, the main data file is copied to a diskette and given to the supervisor, who will then continue processing this file with secondary editing.

Secondary editing

Secondary editing is done with three applications: EDITHH.CN, EDITWM.CN, and EDITCH.CN. The main purpose of these files is to look for and report inconsistencies in the data. These files are discussed in detail below.

EDITHH.CN

This application starts by checking the cluster number. In addition to checking that the cluster number is within valid ranges, it also checks that the region and urban/rural are consistent with the cluster. These will have to be revised based on the sample design in your country. It then checks that the day of interview is possible, for example, it cannot be April 30, nor can it be February 29, 1999 (February 29, 2000 is accepted).

The application then checks that the actual numbers of questionnaires does not exceed the numbers of eligible women and children found in the household questionnaire. Next, it checks that the numbers of eligible household members does not exceed the total number of household members. It then checks the eligibilities for several modules, including child labor and health. It provides an additional check that the household head is at least 15 years of age. This age may be changed to reflect realistic expectations in your country.

The next set of checks relate to education, in questions 16, 20, and 22 of the household listing. These checks examine the grade and number of years at that grade to determine if they are consistent. The limits imposed by these checks are the maximum numbers of years that are spent in preschool, primary school, secondary school, and in a non-standard curriculum. These limits should be modified to reflect the educational system in your country.

EDITWM.CN

This application performs some structural checks. It checks the woman's line number in the household listing as well.

In terms of inconsistencies, this application checks dates of birth and ages of the respondent and her children to determine if they are out of range (children born while the mother was too young, for example), unspecified (neither age nor date given), different in the household listing and the woman's questionnaire, or after the date of interview. Some fields related to ages must be consistent with other information given in the questionnaires, such as the date of the respondent's last birth and the birth of her children in the children's questionnaire.

This questionnaire includes many multiple response questions. EDITWM.CN checks that the responses to these questions are consistent. For questions 2 and 3 in the maternal and newborn health module, it checks

for missing values (all responses must be missing or none must be missing) and whether "No one" was selected and any other option was also selected. Other multiple response questions, including currently family planning method. Antenatal case, and delivery assistance, are all checked for consistency as discussed in the section on MICSWM.CN
.

EDITCH.CN

Edit checking in the children's questionnaire is extensive. It starts by comparing information in the child's questionnaire to the household listing. Error messages will be generated if the child's line number on the cover sheet is the same as the line number for this child's caretaker in the household listing, or if the line numbers for the caretaker in the child's questionnaire is not the same as in the household listing.

The check for the presence of date of birth information illustrates the importance of this data. First, if the month and year are given, they must represent a date that is possible. For example, April 31 is not allowed. Also, the child's month and year of birth or age should be given. If both are given, then they must be consistent.

As in the other editing applications (EDITHH.CN and EDITWM.CN), missing and other values for multiple response questions must be consistent. These variables include drinking of fluids while the child had diarrhea, source of care for illness, symptoms that indicate a child should be taken to a health facility, medicines taken at the health facility, and medicine given for malaria or fever.

Special attention is given to consistency among questions related to vaccinations. If dates are given for vaccinations, several of these are generally given on the same day:

- DPT 1 and polio 1
- DPT 2 and polio 2
- DPT 3 and polio 3
- BCG and polio 0 or polio 1

This may not be the true in your country. If the general practice is to give different vaccinations than these at the same time, this application should be modified to reflect this. In addition, polio 0 should have been given before polio 1, polio 1 should be before polio 2, and so on. The same is true for DPT vaccinations.

If the child's mother reports vaccinations that are not shown on the health card, but the number 66 is not shown in any date of vaccination, this is an inconsistency and the questionnaire should be checked.

As for other dates, the dates of vaccination are checked for being possible dates, including that they were given before the date of interview and after the date of the child's birth.

The last important set of checks is on the anthropometry module. Values entered for height and weight are compared to upper and lower limits that are considered reasonable based on the child's sex and age. Error messages are generated whenever these bounds are exceeded.

Menu systems

There are two menu systems that will be used for data processing, one for data entry, the other for the supervisor. The data entry menu systems is used for entering the main data, entering the verification data, modifying either the main or verification data, and copying the data to diskette. This menu also allows the user to compare the main and verification data, and to make changes in either one, until the two data files are the same.

The supervisor's menu system allows the supervisor to copy data files from diskette, conduct structure checks and secondary edits, back up the data, and export the data to SPSS. Both of these menu systems are described in detail below.

Environment Space

Important: to use these menu systems, you *must* increase the amount of environment space to 4096 bytes. To do this in Windows 95/98, right-click on the icon that will be used to open DOS for IMPS. Left-click on "Properties," then click on the Memory tab. In the "Initial Environment" box, increase the number to 4096.

Data entry menu

Data entry is started using ENTRY.BAT, from DOS or by clicking on a Windows icon. This batch file makes extensive use of DOS environment variables and several utilities developed at Macro. There are many parts of this batch file that you will need to modify based on your specific hardware and software needs.

One of the first lines of this file contains a SETCHECK command:

```
setcheck set s *
```

This statement displays a dialog box which asks the data entry operator for the cluster number to enter. After entering this number and pressing the Enter key, the `setcheck` command will set the value of a DOS environment variable names "s" to the value assigned in the dialog box. This variable is used throughout this batch file to refer to the cluster number[1].

Several lines below the SETCHECK command, the batch file sets several environment variables to establish the locations of programs and data, as G1 through G5. These are based on two environment variables, MAIN and FLOPPY. These two are the only environment variables that should be modified to accommodate the drive letters that are used on your computers. You should not change G1 through G5, since the directory structure should remain unchanged. This is important: you should *not* modify the directory locations of the IMPS applications or data files, only the drive letters.

ENTRY.BAT is started from the DOS prompt as follows:
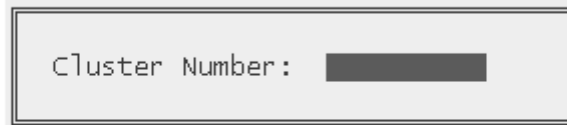
```
Entry <Country_Name>
```

Where "country_name" is the name of your country. This will display the name of your country in the title of the menu when it is shown. It is optional and may be omitted:
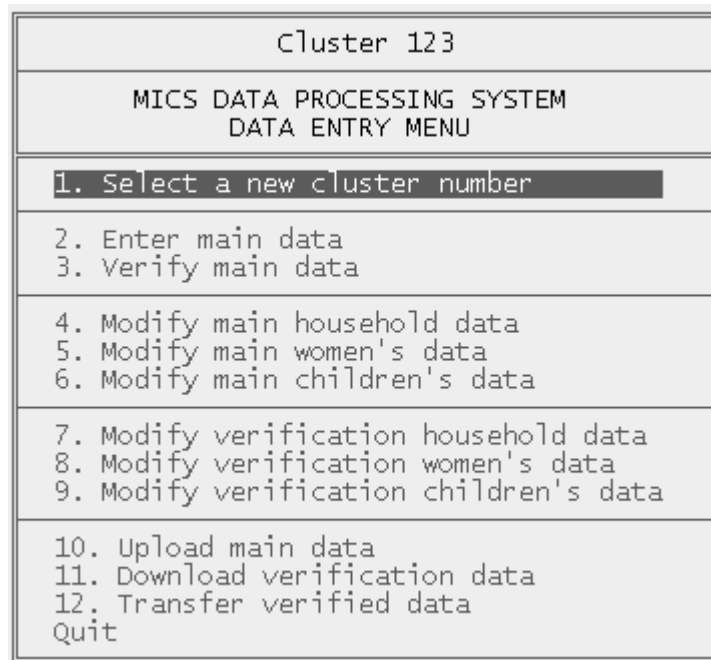
```
Entry Indonesia
```

---

[1] SETCHECK has many uses that are not mentioned here. To get help for this utility, type SETCHECK with no parameters.

Just after starting this batch file, a dialog box will appear asking for the cluster number. This should be the cluster number for the cluster to be entered. This dialog looks like the one shown in Figure 1. Type in the cluster number and press the Enter key.

```
Cluster Number:  ██████████
```

Figure 1: Entering the Cluster Number

```
                Cluster 123

          MICS DATA PROCESSING SYSTEM
              DATA ENTRY MENU

 1. Select a new cluster number

 2. Enter main data
 3. Verify main data

 4. Modify main household data
 5. Modify main women's data
 6. Modify main children's data

 7. Modify verification household data
 8. Modify verification women's data
 9. Modify verification children's data

 10. Upload main data
 11. Download verification data
 12. Transfer verified data
 Quit
```

Figure 2: Data Entry Menu

If the country name is not used the menu will look like the illustration in Figure 2. The options in this menu can be selected by either (1) using the up and down arrow keys to highlight the desired option, then pressing the Enter key, or (2) typing the number or first letter for the desired option. For example, press 7 to modify the verification household data, or Q to quit and return to DOS. Pressing the Esc key is the same as Quit, and will return you to DOS.

There are 5 sections of options available on this menu: select a new cluster, enter or verify main data, modify main data, modify verification data, copy files to or from floppy disk, and quit.

Select a new cluster number

Once this menu is started, every selection that is made relates to the cluster that was selected when it was first started, or until this option is used to select a new cluster. Selecting this option will cause a dialog as shown in Figure 1 to appear, allowing the data entry operator to specify a new cluster number.

Enter main data

Use this option to enter new data into the main data file. Do not use this to verify or change data.

Verify main data

This option is used to re-enter data for the same cluster that was selected earlier. When selected, the data entry operator will enter the same data a second time, but this time in verify mode. This means that each field that is completed will be compared with the original. Whenever there is a difference between the original and new entries, a message will be displayed and the operator will have to enter the correct value. When this process is completed, the revised data file is renamed so it can be distinguished from the original.

Modify main household, women's, and children's data

These options are used only on the computer where the first, or main data was entered.

Modify verification household, women's, and children's data

These options are used only on the computer where the verification occurred.

Upload main data

This option is used only on the computer where the first, or main data entry occurred. It is used to copy the data files to a floppy diskette. This floppy should be given to the supervisor for structure checking. If any changes need to be made as a result of the structure check, they should be done on the same computer as the main data entry. After the structure check is complete, if necessary the data should be uploaded again for use on a different computer (and by a different operator) for verification.

Download verification data

Once the main data entry is complete and uploaded to a diskette, it must be downloaded to a different computer for verification. This option copies the main data (which was 'uploaded' earlier from a different computer) from the floppy to the hard disk drive of the verification computer.

Transfer verification data

This option is used to copy the correct, verified data to a floppy diskette. It should only be used after all differences with the original data file have been resolved.

Quit

Use this option to leave the menu system and return to DOS.

Supervisor's menu

The supervisor's menu is displayed by typing SUPER at the DOS prompt. If the country name is used as a parameter to this command, it will be shown as part of the title of the menu. For example:

```
Super Indonesia
```

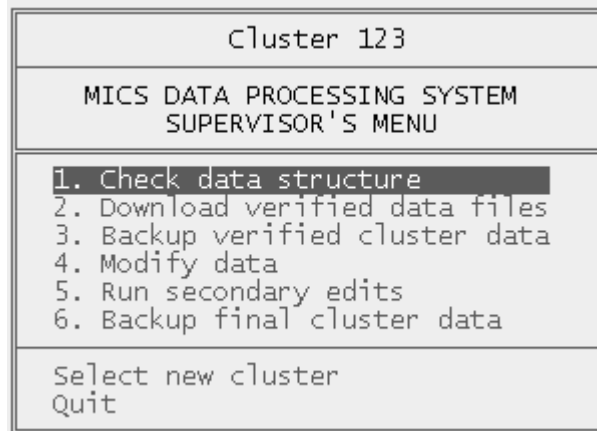An example of a supervisor's menu, without a country name, is shown in Figure 3.

```
┌─────────────────────────────────────────┐
│              Cluster 123                 │
├─────────────────────────────────────────┤
│         MICS DATA PROCESSING SYSTEM      │
│              SUPERVISOR'S MENU           │
├─────────────────────────────────────────┤
│  1. Check data structure                 │
│  2. Download verified data files         │
│  3. Backup verified cluster data         │
│  4. Modify data                          │
│  5. Run secondary edits                  │
│  6. Backup final cluster data            │
│                                          │
│  Select new cluster                      │
│  Quit                                    │
└─────────────────────────────────────────┘
```

Figure 3: Supervisor's menu

Check data structure

This option will run CHECKHH.CN, CHECKWM.CN, and CHECKCH.CN using the data on the floppy disk drive. It will then generate and display reports as described in the section above on checking structures. All of these reports should be printed and carefully examined for structural problems before running secondary edits.

Any errors to the structure will be indicated in the structure report. If any revisions must be made to the data, instruct the data entry operator who entered the main data to make the necessary corrections. After these corrections are made the data should be checked for structure errors again, and as many times as necessary until there are no structure errors in the data. At that point the diskette can be given to the verification data entry operator.

Download verified cluster data

This menu option is used to copy the verified data files from the floppy disk to the supervisor's hard disk drive. It should only be used after the cluster data has been verified and the data entry operator has transferred the verified data to the floppy.

Back up verified cluster data

This option should be used immediately after downloading the verified cluster data. It creates a copy of the un-edited data into another directory.

Modify data

Use this menu option to revise the data that were downloaded earlier.

Run secondary edits

This option will run EDITHH.CN, EDITWM.CN, and EDITCH.CN. All messages that are generated by these applications will be displayed. All of these reports should be printed, and all messages examined carefully to determine necessary revisions to the data. Secondary edits should only be done after running the structure checks.

Backup final cluster data

Select this option only after the structure of the cluster data has been checked and the secondary edits have been run, and all required changes have been made based these reports. This will copy the data files to a backup directory on the same hard disk drive.

Concatenate Data Files

When data entry is complete, three final (i.e., verified, structure checked, and edited) data files will exist for each cluster: one for households, one for women and one for children. These files must be concatenated onto one file each for all households, all women, and all children's data. To do this, at the DOS prompt run the CONCAT batch file by typing:

```
CONCAT
```

This will first create three new files – HH.BCH, CH.BCH and WM.BCH – by concatenating all of the cluster files, then it will copy these three files into the SPSS subdirectory. Note that it is possible to generate these concatenated files before all final cluster data files have been backed up.

These three files can then be read by SPSS, using three SPSS applications that are provided: READHH.SPS, READHL.SPS, READWM.SPS and READCH.SPS (collectively known as the READ programs). These applications will convert the data from IMPS to four SPSS system files, one for households, another for the household listing, a third for women, and a fourth for children.

The four SPSS programs are data description files. They describe the location of case identifiers, record identifiers, variable locations, variable lengths, variable types and names for each variable. Each of these READ programs must be carefully checked and modified if any of the questionnaires used in your country differ from the standard MICS questionnaires. Read the warning in the "Modifying a Read File" section, below.

If you have to modify these READ programs, begin by producing a dictionary list file for each dictionary. These dictionary lists contain all of the information that is needed. When you have created the list files, you must use them to modify the SPSS programs. To illustrate, the following sections examine the parts of the program READHH.SPS that must be checked.

Modifying a Read File

This section uses the READHH.SPS as an example for the changes that must be made to all the READ programs, in order to import MICS data into SPSS.

Data file description

The data file description begins on line 1 with the file type statement and continues until the end file type statement on line 29. Immediately following the file type statement is the keyword grouped. The grouped keyword specifies that the data file to be read contains several records for each case and that some records may be missing or duplicated. The READHH.SPS, READHL.SPS and READCH.SPS programs use the grouped keyword.

The READHL.SPS program uses the mixed keyword. The mixed keyword specifies that the data file contains several types of records that define different cases.

The data file

The first line of the program contains the statement

```
file='hh.bch'
```

This statement specifies the name and path of the data file to be read. If your data file is not in your SPSS working directory, you must specify the full path in addition to the data file name.

Record identifier location

The first line of the program contains the statement

```
record=recid 7-8 (A)
```

This statement specifies the location of the record identifier. In the standard programs, the record identifier is in positions 7 and 8 of the data file and is alphanumeric. SPSS will create a variable named RECID to store this information. Note that at the end of the program the variable RECID is discarded.

To check where the record identifier is stored in your data files, check the first page of the dictionary list. The item name is (record type). Check that the value in the position column is 7-8. If it is not, replace 7-8 in the program with the value in the position column.

The (A) characters specify that the RECID variable is alphanumeric.

Case identifier location

The first line of the program contains the statement

```
case=1-6
```

This statement specifies the location of the case identifiers. In the standard program, the case identifiers are in positions 1 through 6 of the data file.

To check where the case identifiers are stored in your data files, check the first page of the dictionary list. The item names are HI1 and HI2 in the standard dictionary. Check the position column and verify that these two variables occupy positions 1 through 6. If it is not, replace 1-6 in the program with the positions that these variables occupy.

If your questionnaire requires an additional variable to identify each household, your case statement must include the position of this variable. For example, if each household were uniquely identified by HI1 (positions 1-3), HI2 (positions 4-6) and HI2A (positions 7-8), your case statement would be

```
case=1-8
```

Record blocks

Each record type in the data file has a different structure. The variable description information in the SPSS programs is organized in record type blocks. Each block describes the structure of a particular record type.

A record type block is begun by the record type statement and is ended by the next record type or end file type statement. The two-character code contained in quotations marks that follows the record type statement identifies which type of record the block defines. For example, the statement

```
record type "HI".
```

begins the description of the household information record. When SPSS finds a record with the record identifier HI, it will use this record type block to define its structure. The next line contains the statement

```
data list /HI1  1-3
```

This statement begins that description of variables. It creates a variable named HI1 from the characters found in positions 1 through 3 of the record. The lines that follow define the other variables on the record. Variable definition is ended by a period. The location and types of variables are found in the dictionary listing.

Saving the new file

The statement

```
save outfile = 'hh.sav'
  /drop = recid.
```

saves the new file with the name HH and discards the working variable RECID.

The household listing

The READWM.SPS and READCH.SPS programs have the same structure as the READGHH.SPS program. The READHL.SPS program, which creates a household members file, has a different structure. Using the mixed keyword, it reads in household listing records and then saves them in a temporary data file. It then does the same for education module records and child labour records.

When it has created three separate data files, it merges them using cluster number, household number and line number. It then saves this combined data file and erases the three temporary data files.

Warning!

You must check the READ programs painstakingly if you have added or removed any variables. If these programs are not correct, your final results will be wrong. It is easy to make a mistake, but SPSS will only generate error messages in exceptional cases.

It is recommended that the READ programs be checked from start to finish by at least two data processing experts.

Output

When the READ programs are executed, they will produce four SPSS data files: HH.SAV, HL.SAV, WM.SAV and CH.SAV. Each data file corresponds to a unit of analysis.

Creating Analysis Files

Adding labels

Before using a data file, it is helpful to add variable and value labels. A variable label describes the subject matter of a variable. Value labels describe what each value of a variable represents.

Consider the household listing variable HL3 that stores the sex of household members. It takes a value of 1 for males, 2 for females, and 7 for missing data. The variable should be labeled "Sex". The value 1 should be labeled "Male". The value 2 should be labeled "Female". The value 7 should be labeled "Missing".

The SPSS programs LABELHH.SPS, LABELHL.SPS, LABELWM.SPS and LABELCH.SPS (collectively known as the LABEL programs) assign variable and value labels to all the variables in the standard questionnaire. If you have added variables to the questionnaire, go through these programs and add variable and value labels where appropriate.

The LABEL programs are unnecessary if you are using ISSA; ISSA transfers variable and value labels when it exports data.

Adding variables from other files

The SPSS program MAKEHL.SPS, MAKEWM.SPS and MAKECH.SPS (collectively known as the MAKE programs) perform two functions: they add variables from other files and they recode variables to create new variables. Each of these functions will be dealt with in separate sections. This section will concern itself with adding variables from other files.

When analyzing household members, it is useful to have access to variables on the household questionnaire (e.g., HI6, HI7). These variables can be added to the household listing data file by merging it the household data file. More broadly, useful variables can be added to the household listing, women's and children's data files from other MICS data files. The table below shows these three data files and the files with which they are merged.

Base file	External file
HL.SAV	HH.SAV
WM.SAV	HL.SAV
CH.SAV	WM.SAV

These merges are executed in the order implied by tables. As a result, when the household listing file is added to the women's file, it adds both household and household listing variables. Similarly, when the women's file is added to the children's file, it adds household, household listing and women's variables.

SPSS Macros

Macros in SPSS are used to automate a series of statements that are used repeatedly. The MAKE programs use macros to simplify the process of merging files. MAKEWM.SPS and MAKECH use the same macro.

The code below is the macro used in MAKEWM.SPS and MAKECH.SPS.

```
*define a MACRO that merges 2 files with 3 ID variables.
define !merge (file1 = !tokens(1)
              /mvar1 = !tokens(1)
              /mvar2 = !tokens(1)
              /mvar3 = !tokens(1)
              /file2 = !tokens(1)
              /name1 = !tokens(1)
              /name2 = !tokens(1)
              /name3 = !tokens(1)).
get file = !file1.
sort cases by !mvar1 !mvar2 !mvar3.
save outfile = 'tmp.sav'
  /rename = (!mvar1 = !name1) (!mvar2 = !name2) (!mvar3 =
!name3).
get file = !file2.
sort cases by !name1 !name2 !name3.
match files
  /file = *
  /table = 'tmp.sav'
  /by !name1 !name2 !name3.
save outfile = !file2.
get file = !file2.
erase file = 'tmp.sav'.
!enddefine.
```

The define keyword begins the macro and the enddefine keyword ends it. The macro's name, merge, immediately follows the define keyword and is preceded by the exclamation point character.

The text enclosed in parenthesis that follows the macro's name defines accepted parameters. For example, file1 and file2 are parameters that represent file names. When a parameter is referred to in the main body of the macro, it is preceded by the exclamation point. The remaining statements in the macro are standard SPSS commands that will not be discussed here.

The statements above define the macro but do not execute any of the commands within. The commands within are only executed when the macro is called.

A macro is called by typing its name preceded by the exclamation point. When it is called, any parameters used within the macro must be specified. The following lines show the calling of the MERGE macro in the MAKEWM.SPS program.

```
*add variables from the household listing file.
!merge file1=hl.sav mvar1=HI1 mvar2=HI2 mvar3=HL1 file2=wm.sav
name1=WICLNO name2=WIHHNO name3=WILNNO.
```

This statement will merge the files HL.SAV and WM.SAV. The files will be merged using the cluster number, the household number and the woman's line number.

Warning!

The merge above is of a keyed table (e.g., HL.SAV) onto a file (WM.SAV). With this type of merge, extra cases in the keyed table (e.g., household members who are not women aged 15-49) are not added to the file. In other types of merges, these extra cases would be appended to the end of the file. If you modify the match files statement, do not replace the /table specification with a /file specification. This will result in ineligible household members being added to the women's data file. This warning applies to all of the MAKE programs, not just the MAKEWM.SPS program.

Recoding variables

Variables used in the standard tabulation plan are recoded in two places: the MAKE programs and the tabulation programs. Variables that are used in several tabulations are recoded in the MAKE programs. All other variables are created in the tabulation programs.

The recoding of most variables uses standard SPSS commands and will not be discussed here. There are however a few unusual approaches that will be explained.

One of these is the recoding of variables into 0 or 100. This unusual recoding is down for presentation purposes only. When SPSS displays percentages in a table, it displays all categories. For many tables in the tabulation plan, we are only interested in one category. If we assign a value of 100 to that category and a value of 0 to all other categories, the mean of the variable is the percentage of respondents in that category.

For example, the variable HA1 records whether a woman has heard of AIDS. It takes a value of 1 if a woman has heard of AIDS and a value of 2 if she has not. In the first column of table 30, we want to display only the percentage of women who have heard of AIDS.

In the program MAKEWM.SPS, the variable HA1 is recoded into the variable AIDS. AIDS takes a value of 100 if the woman has heard of AIDS and a value of 0 otherwise. The mean of AIDS is the percent of women who have heard of AIDS. To understand why this is so, consider the example below.

Women who have heard of AIDS	10
Total number of women	20
Percentage of women who have heard of AIDS	**10 / 20 * 100 or 50%**
Mean of the variable AIDS	(10*100 + 10*0) / 20
	10 * 100 / 20
	10 / 20 * 100 or 50%

Adding Weights

Sample weights must be used when tabulating MICS data unless your country used a self-weighting sample. There are three kinds of weights: household, women's and children's. Household weights should be used with the household and household listing data files. Women's weights should be used with the women's data file. Children's weights should be used with the children's data file.

Weights must be used if the probability of selecting a household for interview was not equal across all of your country. For example, consider the imaginary country Popsylvania. In the North region, 5 households per 10,000 were selected for interview. In the capital, Vladisville, 28 households per 10,000 were selected for interview. In the country as a whole, 10 households per 10,000 were selected for interview.

Sample weights must be used to adjust the sample to produce accurate estimates for the whole country. The sample weights used are the inverse of the relative probabilities of selection. For example, the sample weights for the North region and Vladisville are

North Region	10/5	2.00
Vladisville	10/28	0.35

The weights for your country are calculated in a spreadsheet named WEIGHTS.XLS. The first step is to run the SPSS program WGTTAB.SPS. This program will produce the information needed to complete the WEIGHTS.XLS spreadsheet. Add this information to WEIGHTS.XLS and it will automatically calculate sample weights in columns O, U and AA.

Save WEIGHTS.XLS when you have made these changes. Then close Excel and open SPSS. From the file menu, select the open option. Select WEIGHTS.XLS and click the open button. SPSS will pop-up the opening files options dialog box. Enter in this box the range of cells in the spreadsheet that you want to read. In the standard version, the range of cells is A5:AD16.

When you have entered the range of cells, click ok. SPSS will read in the specified range of cells and assign variable names to each column based on the values in row 5. In addition to the rows of the spreadsheet that you need, SPSS will also read in three rows that you do not need. Delete rows 1, 6 and 7 from your SPSS data set. Save the file in the SPSS sub-directory under the name WEIGHTS.SAV when these modifications have been made.

WEIGHTS.SPS merges the appropriate weights in WEIGHTS.SAV onto HH.SAV, HL.SAV, WM.SAV and CH.SAV using a macro. WEIGHTS.SPS also deletes unnecessary variables in WEIGHTS.SAV before it merges it with the data files. WEIGHTS.SPS does not need to be modified if your sample is stratified based on area and region.

Output

When labels, variables and weights have been added to these SPSS files and variables have been recoded, you will have four analysis files. The tabulation programs will use these four data files to create the tables in the tabulation plan.

Tabulation

There is one tabulation program for each table. Each program's name is the letter T followed by the number of the table in the tabulation plan. For example, the program T1.SPS creates table 1 in the tabulation plan.

Each tabulation program must be carefully reviewed. One important thing to check is whether the variables used in the tabulation program exist in your data file. If they do not, check whether the variable is of primary or secondary importance. If a variable of primary importance does not exist in your data file, you must either remove the table entirely or ask an analyst to redesign the table. If a variable of secondary importance is missing, remove all reference to the variable in the tabulation program and make any other adjustments necessitated by its absence.

All recoding activity must also be carefully checked. If there are variables on your questionnaire that have non-standard categories, any recoding activity involving those variables must be examined. If your questionnaire contains non-standard variables, they must be recoded if they are to be included in tabulations.

Weighting in the tabulation programs is straightforward except where the SPSS aggregate command is involved. If the goal of the aggregate command is to cumulate across cases to calculate a numerator and a denominator, weights must be applied before the aggregate command. They should not be used when working with the resulting file; it has already been weighted.

For example, table 1 contains the household responses rate. The household response rate is difficult to calculate because it requires dividing one variable by another within the table. One solution to this problem is to create an aggregate file that contains counts of sampled households, occupied households and interviewed households. The aggregate file will contain one case for each category of the specified break variable.

The weights must be applied when the aggregate file is created to generate the weighted numerator (the count of interviewed households) and the weighted denominator (the count of occupied households). Once the aggregate file has been created, the household response rate for each category of the break variable is the numerator divided by the denominator.

If the goal of the aggregate command is to create a summary statistic for individual cases, weights must be applied after the aggregate command. For example, table 4 contains information on the percent of households that contain at least one child under the age of 15.

This information is not present in the household data file, but it can be created by aggregating the household listing file. The break variables are cluster number and household number. Weights are applied after aggregating because we are interested in the weight percent of households with at least one child under the age of 15, not the weighted number of children under the age of 15 in each household.

You must also check any merge operations if your questionnaire uses case identifiers not present in the standard questionnaire. There are a number of merges in the tabulation programs that will only work if unique identifiers are used.

The Include Command

The SPSS program TABLES.SPS can be used to run all of the tabulation programs at once. It consists of a series of INCLUDE commands that execute the tabulation programs individually. If SPSS encounters an error in a program that is included (i.e., executed by an INCLUDE command), it will immediately stop executing the program and return to the program that included the program (i.e., the program that contained the INCLUDE command).

Because of this, you should only use TABLES.SPS when you have checked and modified all of the individual tabulation programs. Be sure also to remove any INCLUDE command that executes a tabulation program that you are not using.

The INCLUDE command imposes four restrictions on the programs that it executes. The first restriction is that each command must begin in the first column of the program. This restriction appears to limit program indentation, but indented lines can be preceded by the '+' character. The commands below, from the program T16.SPS, illustrate the use of the '+' character to maintain program structure.

```
do if (cage >= 6 and cage <= 9).
+ compute solids = 0.
+ if (BF3G = 1) solids = 100.
end if.
variable labels solids "Solid foods".
```

The second restriction imposed by the INCLUDE command is that if a command continues over multiple lines, column 1 of the continuation lines must be blank. Consider the commands below from the program T1.SPS.

```
add files
  /file=*
  /file='tmp6.sav'.
```

Notice that the subcommands on the second and third lines are indented two columns. (While they need only be indented one column to satisfy the restriction, they have been indented two columns to remain consistent with the MICS programming style.)

The third and fourth restrictions are that command terminators are optional and that an asterisk (*) in the first column of a line indicates a comment line. Neither of these restrictions affects our tabulation programs.

In addition to TABLES.SPS, there is an SPSS program that automates the creation of analysis files for each of the three data entry packages. The programs are named EPIINFO.SPS, IMPS.SPS and ISSA.SPS. These programs should only be used when all of the component programs have been executed and shown to work.

These programs are useful for recreating analysis files when a change is made to one of the file creation programs. They insure that all of the programs will be executed and that they will be executed in the proper order.