Process documentation - Data Management
IBBA Round I
(Draft)

A project document

Indian Council of Medical Research (ICMR)
& Family Health International (FHI)

Contents

Background and Overview

1.0 Preparation

- *1.1 Designing Questionnaire*
- *1.2 Generating Identification Number*
- *1.3 Setting up of Data Processing Team and place*
- *1.4 Setting up a system for managing Questionnaires and Data files*
- *1.5 Developing and Testing Computer Programs*
- *1.6 Training of the Data Processing Team for various activities*

2.0 Primary Data Processing

- *2.1 Office Editing of the Questionnaires received from the field*
- *2.2 First level Data Entry (Behavioral data)*
- *2.3 Second level Data Entry (Behavioral data)*
- *2.4 Comparison, Verification, Modification and Cleaning of double entered data set (Behavioral data)*
- *2.5 Final Editing and Cleaning Data*
- *2.6 Biological Data Entry*
- *2.7 Coupon Entry for RDS groups*

3.0 Secondary Data Processing

- *3.1 Exporting data into SPSS*
- *3.2 Secondary Editing of Clean data files*
- *3.3 Merging of Behavioral and Biological data files of district specific groups*
- *3.4 Calculation of Sampling Weights*
- *3.5 Backup of Final Merged Data Files*
- *3.6 Coding and Recoding of variables*
- *3.7 Data Analysis*

Background and Overview

Integrated Behavioural and Biological Assessment (IBBA) is a multi-centric study consisting of 64 groups in twenty nine districts across six high HIV prevalent states and along four selected route categories of the National Highway of India. The different high risk groups covered Female Sex Worker (FSW), Men having sex with men (MSM), Clients of FSW (Clients), Injecting drug users (IDUs), Truckers and Hijras. The sampling methodologies, data collection procedure, the questionnaire, the data entry formats were different for different groups across the districts and states. Also, it was a study where both behavioural and biological data were collected. Data collection and data management of IBBA was a major challenge and played a major role in this study.

The data processing system in IBBA can be divided into three phases:

(1) Preparation
(2) Primary data processing and
(3) Secondary data processing

The preparatory work needs to be done to ensure the smooth flow of questionnaire from the field, setting up of trained data management team and being ready for uninterrupted data entry. **Preparation** stage comprised the following steps:

(1) Ensure proper coding, skip patterns, layout, structure, numbering, formatting etc. from the data management point of view in the questionnaire and suggest for modification and changes before the questionnaire is finalized and goes to the field for survey
(2) Generating identification of respondents (ID Number)
(3) Setting up a data processing team and the place
(4) Setting up a system for managing the questionnaires and data files
(5) Developing and testing computer programs developed in CSPro and Excel for the district specific questionnaires and laboratory formats
(6) Training for the data processing team for various activities

The goal of **Primary data processing** was to produce clean, edited data files. It comprised of the following steps:

(1) Office editing of questionnaires received from the field
(2) First time entry of all questionnaires for a district specific group into a data file
(3) Second time entry of the same data into a different data file
(4) Compare, verify, modify and clean the data sets based on two level of entries
(5) Backing up the checked and verified data file
(6) Secondary editing of the clean data file
(7) Backing up the edited, or final, data file

The steps (4-7) above were iterative process which is repeated until all problems had been resolved or all remaining issues had been determined to be acceptable.

The secondary data processing was to analyze data and produce output based on the analysis plan. This comprised the following steps:

(1) Exporting the data to SPSS
(2) Secondary editing of the clean data file
(3) Merging district specific behavioural and biological data files
(4) Calculating sample weights and insert into the data file
(5) Backing up the edited, or final, data file
(6) Recoding variables to simplify analysis
(7) Generating top line finding
(8) Creating the descriptive tables required
(9) Creating the exhaustive tables required
(10) Further statistical analysis
(11) Backing up syntax and outputs

1.0 Preparation

1.1 Designing Questionnaire

At the time of designing and development of the questionnaires, involvement of data management personnel is very important. The data management personnel had put his/her views and suggestions for any modification or changes related to layout, structure, question numbering, skipping patterns, multiple answer response pre-coded responses, in the questionnaire. For example, the pre-coding pattern of 'Not applicable' (96, 996, 9996), 'other' (97, 997, 9997), 'Don't know' (98, 998, 9998) 'No answer' (99, 999, 9999) categories depending on number of digits were uniform across all the questions in the questionnaire. The questionnaires were checked in all respect so that all the valid answers are captured by the data entry program. Standardization of questionnaires across districts and groups were proposed but since the survey was done in phases, the questionnaires got modified time to time resulting in more than one version.

1.2 Generating Identification Number

Since there were many groups under each domain (FSW, MSM, IDU etc.), it was necessary to develop systematic unique identification number (ID No.) for the respondents. Fifteen sets of stickers with unique ID numbers were pre-printed for the identification of the respondent to be used for behavioural questionnaire and different biological tests to cover 400 samples per group.

The identification number consisted of seven digits – first two digits represented the district code (Mumbai, Chitoor, Karimnagar etc.), third and fourth digit represented the domain code (FSW, MSM, IDU etc.) and last three digits represented the respondent number within the specific district and domain. For example, in ID number 2402157, '24'

is the code for the district Pune, '02' is the brothel based FSW group and '157' is the respondent number of Pune brothel based FSW group.

1.3 Setting up of Data Processing Team and place

Four data processing teams were established: one at the central level and the other three at the state level. At the state level, the research agency and the state ICMR institute - both had separate team comprised of a data manager and 3-4 data entry operators. The data processing team at the central level called DMG (Data Management Group) was set up with four people with strong data management and analysis skills, and statistical background. The technical support and assistance to all the teams was provided by FHI.

1.4 Setting up a system for managing Questionnaires and Data files

The data processing unit was a separate unit with well configured computers for the data entry operators and the data manager, a printer, secondary data storage device, UPS and questionnaire storage space.

1.5 Developing and Testing Computer Programs

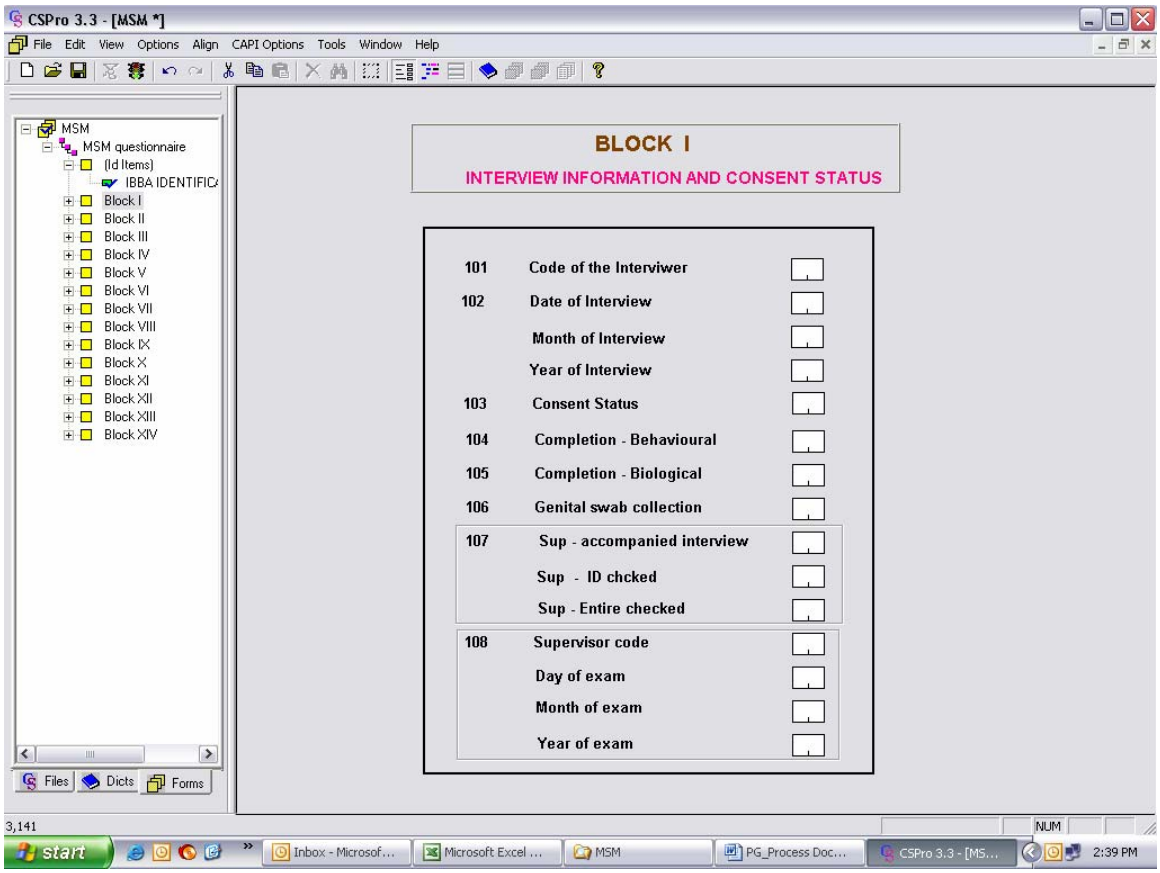
Development of the data entry program was designed and developed by FHI for each individual district specific groups which were tried out at the state levels for the modification and changes required for debugging. CSPro (Ver 3.0) was primarily used for the data entry of behavioral questionnaires and Excel was used for biological data entry.

CSPro (Census and Survey Processing System) is a public-domain software package for entering, editing, tabulating and mapping census and survey data. CSPro was designed and implemented through a joint effort among the developers of IMPS and ISSA: the United States Census Bureau, Macro International, and Serpro, S.A. Funding for the development is provided by the Office of Population of the United States Agency for International Development. CSPro is designed to eventually replace both IMPS and ISSA. CSPro combines and expands upon the capabilities of both ISSA and IMPS. It takes advantage of the power and flexibility of both of these programs, but adds the friendliness, ease of use, and intuitive nature of Windows. CSPro provides a more visual approach to the creation and manipulation of data and reduces the programming needs. This easy to use package facilitates defining data structures, developing applications, entering and checking data and generating reports. More advanced users, including computer programmers, can access the full CSPro language to perform complicated tasks. Excel was used as the data entry program for entering biological data with proper consistency and range checks. Also coupon entry was carried out in Excel for the RDS groups.

In CSPro, dictionaries are used to describe the data structure: a group of related variables (questions) comprises a record (module), and a group of records comprises a level (questionnaire). These are stored in a dictionary file (extension: *dcf*). In addition to the

data dictionary, forms linked to the dictionary are used for data entry. There is usually one form for each record. The forms are stored in a forms file (extension: *fmf*). The *dcf* and *fmf* files can be modified directly. The best way to do this is to open the forms file in CSPro. This will give access to the data dictionary and the forms together and ensure that the two remain synchronized. Dictionary is the data structure and the forms are the visual replication of the variables defined in the dictionary on the screen. After the dictionary and the form files are finalized, an application file (extension: *app*) is created where all the logical checks are taken care of. Consistency checks, range checks and automatic skip wherever applicable were built-in while developing the group specific data entry programs. The error message was also displayed during data entry in case the data entered for a particular question is inconsistent or it is out of range. Thus, after the data entry program is complete, it is tested by entering dummy data to find errors, if any, and according debugging of errors was carried out.

A typical CSPro data entry screen looks like the following:



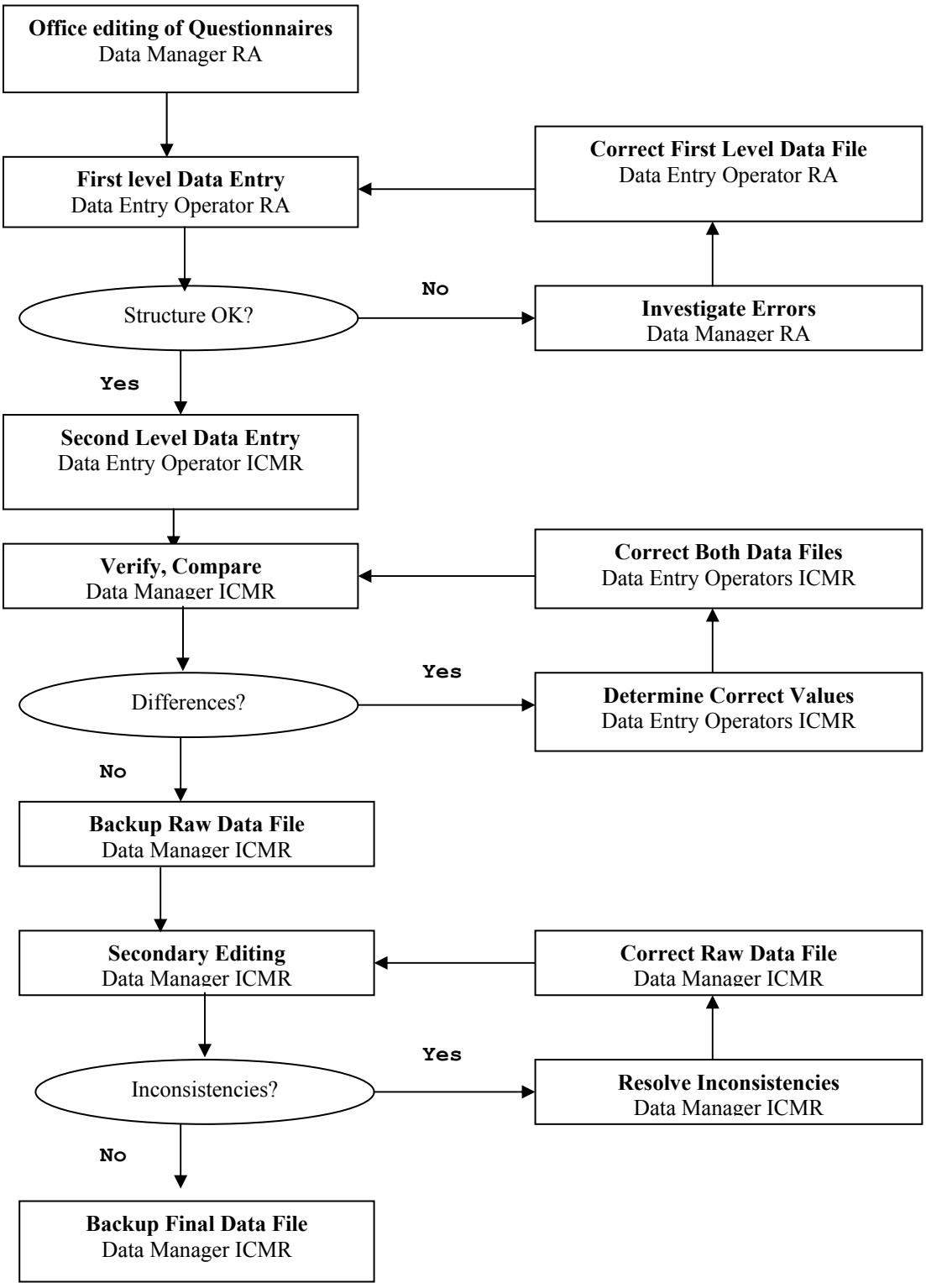
1.6 Training of the Data Processing Team for various activities

Training was given to the data processing teams on the management of questionnaire which comes from the field, office editing of the questionnaire, post coding of open

ended questions, data entry program, backing up data and other related aspects of data management.

2.0 Primary Data Processing

Flow chart



2.1 Office Editing of the Questionnaires received from the field

The data managers responsible for data entry and data cleaning must have an excellent understanding of the questionnaires and the goals of the survey and was present in the interviewers' training session. Interviewers' manuals were provided to him/her to support understanding inconsistencies. Proper training was given to him/her on the software used for data entry. Another responsibility of the data manager was to do the post coding for the open ended question answers and simultaneously keep record of those post coding in a separate excel file.

When the questionnaires for a district specific group arrived from the field, the data manager had to check the number of questionnaires against the control sheet. If any questionnaire was missing, the data manager had to contact the fieldwork team and see if the questionnaires could be found. It is advisable to arrange the questionnaires in ascending order of the ID number. This would help maintaining the questionnaire and also the data entry easier. The next task of the data manager was to go through each of the filled-in questionnaire – check and verify that all the relevant questions were filled up and legible. Editing of the filled-in questionnaire in terms of inconsistencies are very important because the data entry operators would be stuck up if the inconsistencies in the answers given by the respondents are not sorted out before hand. An example for a consistency check - if the current age of an FSW is entered as 25 years and the age at first sex is entered as anything more than 25 years, the program will show an error message as "Age at first sex cannot be more than the current age…… please check". The data manager also had to investigate and resolve such inconsistencies, some of which could be complex, in the filled-in questionnaires and should edit wherever necessary before sending these for data entry. It happened that some of the inconsistencies could be sorted out by the data manager. In that case he had to take help from the field supervisor. All of the editing on the questionnaire was done by a different color pen by the data manager. The data manager also transferred all the responses to the code boxes given on the right side of the questionnaires which in turn helped data entry operators at the time of data entry.

2.2 First level Data Entry (Behavioral data)

The data entry operators' job was to enter the data into the program. They had prior data entry experience and were made familiar with the questionnaires. Before beginning data entry, a separate two or three day training was held to make the data entry operators familiar with the data entry program and the data processing system. By the end of the training, the data entry operators were comfortable with the data entry program and became aware of their daily responsibilities under the supervision of the data manager. The data entry operators were not supposed to apply their own logic in case of any inconsistencies encountered during data entry and the data manager had to sort out the issue before they could proceed with the data entry.

Because data are saved only after all the relevant questions in a particular questionnaire have been entered, data entry operators should not leave their computers in the middle of

entering data for a particular questionnaire. Before taking a break or stopping work for the day, all of the responses given by a respondent in the specific questionnaire had to be completely entered. Further, it was recommended to copy the data onto the data manager's computer or a pen drive as a precautionary measure before leaving for the day. In addition, every evening the data manager had to copy the contents of the data entered in different computers by different data entry operators onto the secondary storage device. This safeguard allowed the data manager to recover if any of the computer crashes.

In addition to controlling which questionnaires had been entered, the data entry application rigorously controls the skip pattern within a questionnaire. That is, it asked for the responses to questions that should have been asked given the responses to previous questions. For example, in the HIV/AIDS knowledge section, if the answer for the very first question (say, whether the respondent ever heard of HIV/AIDS) is 'No' it skips to the next section. The data entry application also designed in such a way that the cursor automatically goes to the next section for that particular answer.

The second task of the data entry application was to minimize data entry errors. The data entry application does this by performing checks as the data were entered. If a value entered for a question was outside the range of values on the questionnaire or if some other basic inconsistency was detected, the data entry application displayed an error message and required the data entry operator to resolve the inconsistency in consultation with the data manager before advancing forward. More complex inconsistencies, whose resolution would slow down data entry considerably, were not checked during the data entry process and those were checked instead during secondary editing.

After the data entry was over for a survey group, the data manager had to check and verify the data entered with respect to the questionnaires in hand. Then he had to merge all the data files entered by the operators into a single data file. The set of questionnaires, the data entry program, the data file and the list of post coding were then submitted to the state ICMR institute for second level data entry and verification and cleaning of data.

2.3 Second level Data Entry (Behavioral data)

The second level data entry was carried out by the state ICMR institutes after they received the questionnaires, the first level data files and the code list from the research agencies. The data processing team in the state ICMR institute also comprised of a data manager, 3-4 data entry operators. The responsibilities of these personnel are a bit different from those at the research agencies though there are some common activities.

The data manager, while receiving the set of questionnaires, the data files and the code list from the research agencies, had to ensure all are in order. At initial stage the data manager prepared for second level data entry and distributed the questionnaires to the data entry operators. The data entry procedure started immediately under the supervision of the data manager as described earlier in case of research agencies. The data manager checked and verified that all the questionnaires were entered properly and merged into a

single file as said earlier. In case of any inconsistencies found during the second level entry, the data manager had to clarify from the research agencies. There were other activities to be performed by the ICMR data processing team after the second level data entry was over, which are detailed below.

2.4 Comparison, Verification, Modification and Cleaning of double entered data sets (Behavioral data)

Comparison of the data files entered by the research agency and the ICMR institute was the next step. This option compared the two files using CSPro's comparison tool. If there were any differences between the data files those were displayed on the screen. This output had to be printed and given to the data entry operators responsible for entering the data. The data entry operators then consulted the questionnaires and determined the correct value for each instance in which their data files disagreed. Once they had determined the correct values, the operators had to update the respective data file. This was an iterative process and at this point the files were compared again and again. When no differences between the two files remained, the data file could be finalized. It is to be noted that while there were no errors left, both the first and second level entry became basically the same file.

For example, let us suppose first level entry of a particular group has been entered into the file A and the second level entry of the same group is entered into the file B. The comparison was done from both ends, i.e A was compared with B and vice versa. Say, for a particular question in A the code entered was '2' while for the same question in B it was entered as '3' which would be flagged in the error list while comparing A and B. Then looking at the ID one had to physically check the corresponding questionnaire and correct the entry in A or B wherever applicable. Similar process was done for all the flagged variables in the error list. Both way comparisons were necessary because one way comparison A to B would only check for IDs of A. So, if entry of any record is missed out in A would not be captured in the error list. At the end when there were no errors shown while comparing A and B, both the files becomes basically identical files. But the original files A and B before comparing were kept in a separate folder for record. The error list generated looks like below:

Input File: Pathname\A.dat Reference File: Pathname\B.dat

Case Id/ Item	Input File	Reference File
[0311401]		
TE code	347	847
TLC/PSU Code	987	981
Q301	34	43
[0311419]		
TLC/PSU Code	324	234
Q402	5	6
Q515B2	6	7

In the above error list the operators needed to check and verify from the questionnaire referring to ID [……..], which of the entries in the input file (A.dat) and the reference file (B.dat) are correct and accordingly modify the entry in the corresponding data file. For example, after finding the questionnaire with ID [0311401], the operator had to check whether the 'TE code' is 347 or 847. If 347 is the actual data shown in that questionnaire then the reference file (B.dat) was to be corrected for that ID. Similar procedure was carried out for all other listed IDs and the corresponding variables.

2.5 Final Editing and Cleaning Data (Behavioral data)

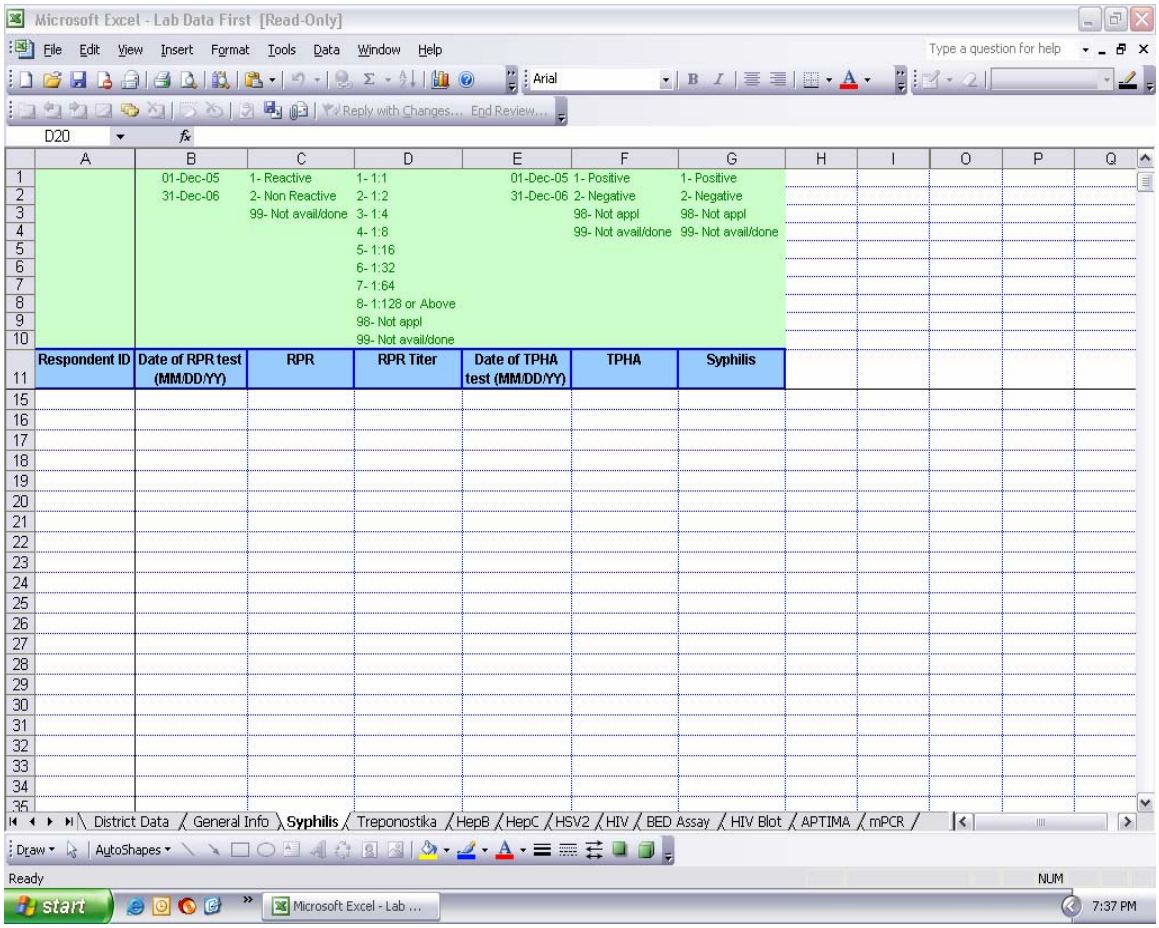
Once this iterative process is over and the comparative list stopped giving any error, any of the input file or the reference file could be saved as the final data file. All the hard copies of the error lists during this iterative process were stored properly for future reference. Then the final data file along with the code list for open ended questions were sent to the central level ICMR data management cell called Data Management Group (DMG) for further secondary cleaning and editing of data before analysis began.

2.6 Biological Data Entry

There were few biological indicators (variables) to be entered compared to behavioral variables. Those were HIV sero-positivity, Neisseria gonorrhoeae (NG), Chlamydia trachomatis (CT), Syphilis test comprising of Rapid Plasma Reaign (RPR) test and Treponema pallidum Haemagglutination Assay (TPHA), Herpes simplex virus (HSV-2) (10% cases), Hepatitis-B and Hepatitis-C.

In the same way as in case of behavioral data entry, double data entry was carried out in biological data entry. But this was done in Excel worksheet. As there were few variables and the results for all the bio-indicators were not available at one go, CSPro could not be used as the data entry software. Excel was a better option for biological data entry of course with consistency and range checks. Separate spreadsheets were used for individual bio indicators with respondent ID as the key variable. The research agencies were not involved in the biological data entry process. Both first level and the second level entries were carried out at ICMR and the discrepancies were sorted out at ICMR before finalizing.

ICMR received the biological test results in paper format from the laboratories in phases for different biological tests for different district specific groups. It was ICMR's responsibility to ensure all tested results were entered into the respective data files twice and also clean the data in case of inconsistencies as per the second level data entry. The data entry format was in Excel with twelve different sheets for twelve types of biological test results entered for all the respondents who agreed for biological tests. The structure and the format for entering biological data looks like the following.



The same format was used for second level entry which included macros for verifying and checking the first level entry. Following are the important points to be noted while entering biological data.

- Data were entered test-wise for each district specific group. Every district specific group had two MS Excel files - One file (Lab Data First.xls) for first data entry and other one (Lab Data Second.xls) for second data entry.
- These excel files had 12 sheets containing district info, group info, sample info and results of each test.
- Data was first entered in Lab Data First.xls file and then in Lab Data Second.xls file.
- District Info sheet in the file was filled at the time of starting data entry and range of IDs were entered correctly.
- Validation ranges for data had been shown in the top rows of each field/column. Only valid data could be entered.
- Second data entry also was done in same sequence of IDs as it was done in first entry.

- While doing second data entry it had to match data with data in first entry. In case of data mismatch it showed red color in that cell alerting for punching error.
- Data entry operator rechecked the data in case the cell became red and kept a note of it and made correction in the first or second file depending on the correct entry after completing the second data entry.
- For efficient file management, files had following directory structure State→ District → Group→ First and second data entry files.

2.7 Coupon Entry for RDS groups

For all the groups where respondent driven sampling (RDS) was adopted, there was another activity of data entry to enter the coupon numbers received by the respondent and the coupon given to the respondent for further distribution. The seeds who were selected for interview, were given three coupons each for distribution. The person who came for interview with one of these coupons was again given another set of three coupons for further distribution. The coupon with which the respondent came for interview is called the 'primary' coupon and the three coupons given to the respondent for distribution is called 'secondary' coupon. RDS analysis tool (RDSAT) does not run without the primary and the secondary coupon numbers entered in the data file against each of the respondent. Excel was used for entering these coupon numbers. The coupon data entry format looks like the following

Resp ID	Primary	Secondary_1	Secondary_2	Secondary_3
3110001	1	11	12	13
3110002	2	21	22	23
3110003	22	221	222	223
3110004	11	111	112	113
3110005	8	81	82	83
3110006	82	821	822	823
3110007	81	811	812	813
3110008	83	831	832	833
3110009	21	211	212	213
3110010	13	131	132	133
3110011	12	121	122	123
3110012	112	1121	1122	1123

In the above format, the first column is the unique ID of the respondent, the second column is the primary coupon number, and next three columns are the secondary coupons. For example, ID 311005 who is the seed no.8 was given secondary coupons 81, 82 and 83. The person who came for interview with say, coupon number 83 was given three coupons numbered as 831, 832 and 833.

In entering coupon numbers like above, there were chances of making mistakes as the numbers are mostly combinations of 1s, 2s and 3s and the number of digit increases and

the chain of respondents increase with a particular seed. So, the data cleaning is very important before this is merged with behavioral and the biological data. Manual check becomes impossible. Checking of these number were done by applying formula in the excel sheet which checks (1) a particular primary coupon number should have presence within the set of secondary coupons and that also only once, (2) There should not be any duplicate coupon number across primary and secondary coupons, (3) The secondary coupon numbers can appear in the list of primary coupon numbers either once or not at all.

3.0 Secondary Data Processing

3.1 Exporting data into SPSS

When primary data processing of behavioural data was complete, a clean data file for each district specific group became ready. While primary data processing was done using CSPro, secondary data processing was done primarily in SPSS Version 14.0. The first step in secondary data processing was therefore converting the data from CSPro's data format to SPSS' data format. This was done using the "Export the data to SPSS" option CSPro Tool menu.

When this option is selected, the data file is then exported to SPSS by the *export.bch* application. This application creates an ASCII data file and a syntax file. While the SPSS data description files will read the ASCII data files into SPSS, they do not save them. To get the data description files to save the data in SPSS format, the SPSS command

save outfile = `filename.sav'.

must be added to the end of each data description (syntax) file. The word 'filename' should be replaced by a name depending upon the type of data file. Once this command has been suitably modified and added to each data description file, executing the SPSS data description files will create the SPSS data files *filename.sav.*

Simultaneously the biological data sets and the coupon numbering data sets wherever applicable were also converted into separate SPSS files. Biological data sets were converted into a number of individual SPSS data files depending on the indicators applicable for the district specific group (HIV test, NG, CT….. etc.). All the individual data sets had the common variable ID as the key variable.

3.2 Secondary Editing of Clean data files

Frequency checks were carried out for each and individual group data sets to find other inconsistencies like outliers, denominator or base for each valid response etc. in the data which could not be verified during data entry. Any inconsistencies were resolved after discussion with a team of technical people before proceeding for data analysis.

3.3 Merging of Behavioral and Biological data files of district specific groups

This is a very important step to merge the behavioural and the biological data files for a district specific group (also RDS coupon numbering wherever applicable) in SPSS and precautions had be taken that the merging was done properly. The steps followed for merging is given below:

The SPSS file for HIV test results as told earlier was taken as the master file for merging. All the records which had the HIV test results were kept in tact and the records which did not show any result (due to various reasons) were removed from the file. Then the file was sorted based on ID number and saved.

All other individual SPSS files to be merged were also sorted on the ID variable and saved. Considering HIV file as the master file all other biological indicator files and the behavioural files were merged together for a district specific group, ID taken as the key variable.

After the biological and the behavioural data were merged together, it was necessary to verify if merging had been done successfully by checking individual data records in the merged file and comparing those with individual SPSS files.

3.4 Calculation of Sampling Weights

The sample weights were calculated and included in the final set of data.

Assumptions:

- Clusters are selected by systematic PPS sampling (Measures of sizes of clusters, M_i and cumulative measure of sizes of cluster universe, M are the inputs).
- Samples from each selected clusters were drawn by Simple Random Sample.
- Since re-sampled clusters were few, they were considered as selected from the primary selection.

Weight calculation was not necessary in the groups where Respondent Driven sampling method was adopted or the groups where 'Take All' strategy was adopted.

Except Mumbai (Street based and Brothel based FSW), the sample selection are by two stage sampling. Mumbai has segment selection as first stage in the three stage sampling.

.

Procedure for Weight calculation in non-RDS groups

Obtain total measure of sizes (CMoS= M) of all the clusters (sampling universe) considered for sampling. Separate CMoS should be obtained from Sampling Frame (SF) for TLC and Conventional if the selection procedure involves both. Obtain measure of sizes (MoSi= M_i) for each selected cluster from Sampling Frame. From the selected

clusters; obtain eligible respondents (N_i) ; eligible respondents selected for interview (a_i) ; and number of respondents completed (both biological and behavioral) (n_i), available in the Cluster Information Sheet (CIS). Also get the selected cluster IDs and cluster type (Time Location Clusters - TLC or Conventional - C). These IDs (unique in all respect) should have linkage to other information that is available in the SF.

a. Check if $N_i \geq a_i$. If this condition is violated, a_i is equated to N_i.

b. Check if n_i obtained from data sets (from completed proforma) is equal to that obtained from CIS. If the difference is high, then ask for clarification for possible coding errors. After clarification, for weight calculation the 'n_i' is fixed.

c. Check if $a_i \geq n_i$ .If this condition is violated a_i is equated to n_i.

d. If $n_i = 0$, then that cluster is not at all used for weight calculation.

NB: The corrections suggested above should be made only if the discrepancies are within acceptable level. Other wise best possible estimates are made from the available information. For example, 'a_i's may be estimated from the mean ratio of 'n_i' to 'a_i' of available data.

A separate data sheet is created for Cluster type, cluster ID, MoSi, N_i, n_i, and a_i before start doing the calculation.A typical format of the data sheet for weight calculation is given below.

Type	Cluster ID	MoSi=measure of size for cluster i(M_i)	Total no. of eligible resp. in cluster i (N_i)	Number selected for interview (a_i)	Number completed (Beh.+Bio) in cluster I (n_i)
c	1	45	40	14	12
c	2	38	36	12	9
c	3	28	24	11	9
c	4	28	25	11	9
c	5	48	42	14	12
c	6	38	33	12	9
c	7	48	44	18	12
c	8	48	42	14	11
c	9	28	26	10	7
c	10	33	31	12	7
c	11	28	26	8	6
c	12	28	25	9	7
TLC	13	4	3	3	1
TLC	14	28	22	14	9
TLC	15	13	10	5	3
c	16	45	42	15	12

TLC	17	33	28	15	12
TLC	18	33	31	14	12
TLC	19	33	29	13	9
TLC	20	28	24	10	9
TLC	21	8	6	3	3
c	22	23	22	7	6
c	23	28	26	10	9
c	24	33	27	11	9

The formulae and procedure.

The inclusion probability of a site/cluster (PPS) and a sub sample selection of a_i individuals (Simple Random Sample) from N_i of the ith cluster is obtained through

$$P_i = (m \times \frac{M_i}{M} \times \frac{a_i}{N_i}) \qquad (1)$$

where m is the number of clusters to be chosen for data collection

M is total measure of size of the survey universe(eg. Total number of documented FSWs in a district)

M_i is the measure of size of the ith cluster

N_i is the estimated size eligible in the ith cluster

a_i is the number selected for interview

(M_i may be more; or less than N_i)

Note:

i. The m and M values correspond to type of clusters(TLC or Conventional)

For example in Andhra Pradesh , Chitoor- FSW; 73 clusters were selected, of which 20 were TLCs . When inclusion probability for TLC is calculated ' m ' will become 20 and the corresponding M will be 1363, the total size of the universe for TLC population in Chitoor district. And the corresponding 'm' for conventional clusters will be 52 with appropriate 'M'.

ii. ' n_i ' is not used in (1)

iii. Formula (1) is modified accordingly if more than two stages are involved in the sample selection. For example in Mumbai, there were three stages of selection. Stage I was selection of segments; say C out of S segments, and if the selections are by Simple Random , then the equation (1) becomes

$$P_i = (m \times \frac{M_i}{M} \times \frac{a_i}{N_i}) \times \frac{C}{S}$$

NB: It is possible that some P_i >1 , when M_i are greater that the sampling interval. If such events are very few the P_i were equated to '1' implying that the larger clusters were selected with certainty

.

Sampling Weights:

The sampling weights for the individuals in the i[th] cluster is given by

$$w_i = 1/P_i$$

P_i - is the inclusion probability of selection for individuals in the i[th] cluster

Standardized weights.

When the sampling weights are attached for weighted estimates, the number of observations will thereby be altered especially be inflated, resulting different sample size than was actually realized in the survey. This necessitates a need for a correction. This is done by standardizing the weights, standardized for the total sample size (In our case it is separately done for conventional and TLC).

The Standardized weight for i[th] cluster is obtained by using the general formula

$$w_i' = \frac{w_i \times \sum n_i}{\sum (w_i \times n_i)}$$

This formula for the conventional cluster is

$$w^C{}_i' = \frac{w^C{}_i \times \sum n^C{}_i}{\sum (w^C{}_i \times n^C{}_i)}$$

This formula for the time location cluster is

$$w^T{}_i' = \frac{w^T{}_i \times \sum n^T{}_i}{\sum (w^T{}_i \times n^T{}_i)}$$

T and C represent values corresponding to time location and conventional clusters.

Adjustments in case of incomplete (Take all) data

The following adjustments were made in the calculation of weights if few of the input data were incomplete (Take all).

- If N_i was not available , corresponding MoSi (M_i) was used and vice versa for N_i .

- If a_i was not available, response rate, $\frac{n_i}{a_i}$ was calculated from the available data and a_i was estimated by using n_i and the average response rate. If the estimated a_i was more than N_i then the value of a_i was equated to N_i .

- If MoSi , N_i and a_i were not available for a cluster then all these three values are equated to n_i .

Cumulative measure of size of the universe was not changed.

A spread sheet is given below with self explanatory columns and formulae. Type of cluster, cluster Id, M_i , N_i, n_i ,and a_i are copied from the data sheet referred in the previous table and pasted accordingly into the spread sheet. CMoS values for TLC and Conventional clusters are typed into the appropriate cells.

Type of cluster	No. of Clusters selected (m)	CMOS by type of cluster.	No.Completed (Beh+Bio) by type	cum Wi*ni by type				for adjusted wt find the response reatio ni/ai		Adjusted Pi		ADJ Standardized weight	Wi=weight for ith cluster	
- Conv	53	1240	278	926.3										
- TLC	20	1363	123	985.8										
Type	Cluster no	MoSi=expected measure of size for cluster I	ni = no of completed Behav+Bio intvw in cluster i	Ni = total no of eligible resp in cluster i	PM=Mi/M	PN= ni/Ni	Pi=m*PM*PN	No. selected for interview (ai)	response rate (RR)	Adj pi=Pi/RR	ADJ Wi = 1/pi	Wi'	Wi*ni	Crosscheck
C	1	9	6	9	0.007	0.667	0.256	8	0.750	0.342	2.925	0.878	17.547	5.266
C	2	7.5	5	7	0.006	0.714	0.229	5	1.000	0.229	4.367	1.311	21.836	6.553
C	3	8.5	5	8	0.007	0.625	0.227	6	0.833	0.272	3.670	1.101	18.350	5.507
C	4	12.5	9	13	0.010	0.692	0.370	12	0.750	0.493	2.028	0.609	18.249	5.477
C	5	4.5	3	5	0.004	0.600	0.115	4	0.750	0.154	6.499	1.950	19.497	5.851
C	46	12.5	9	15	0.010	0.600	0.321	10	0.900	0.356	2.808	0.843	25.268	7.583
TLC	49	7.5	4	6	0.006	0.667	0.073	6	0.667	0.110	9.087	1.134	36.347	4.535
TLC	50	22.5	9	20	0.017	0.450	0.149	11	0.818	0.182	5.507	0.687	49.564	6.184
TLC	51	22.5	9	21	0.017	0.429	0.141	10	0.900	0.157	6.361	0.794	57.246	7.143
TLC	52	12.5	6	12	0.009	0.500	0.092	7	0.857	0.107	9.346	1.166	56.078	6.997

Wi' is the standardized weight for the respective clusters. When these weights are applied during data analysis and generation of tables, it does not inflate the sample size and remains the same. The last two columns are only for cross check and verification of the sample size. While applying weight, it is always advisable to take maximum number of digits (5-6) after the decimal point for accuracy of the sample size to be maintained. It

can be noted that the mean of the sample weights when all respondents are considered is 1. The standardized weights are inserted in the respective data files by generating syntax in SPSS and saved as the final file for data analysis.

3.5 Backup of Final Merged Data Files

At this stage all the steps before starting data analysis becomes complete. At the ICMR central level, district specific data for the respective groups were saved in a folder and a backup of the same was maintained separately.

3.6 Coding and Recoding of variables

The structure of the data file during primary data processing simplifies the process of entering the data. This structure is not optimal for analyzing the collected data however, so the first task after the data have been transferred to SPSS is recoding variables to make analysis easier and more efficient. This task is known as creating new variables, renaming variables categorise variables in the data file based on the analysis plan. For example, we can recode a continuous variable 'Current age' into a new variable 'Age_cat' with three categories as 18-21 yeas, 22-25 years and 25 & above. We can also combine two or more variables into a single variable where it makes sense. These are iterative process depending on the analysis plans. These steps were carried out for all the individual data sets at different phases depending on the analysis plan and saved as SPSS syntax.

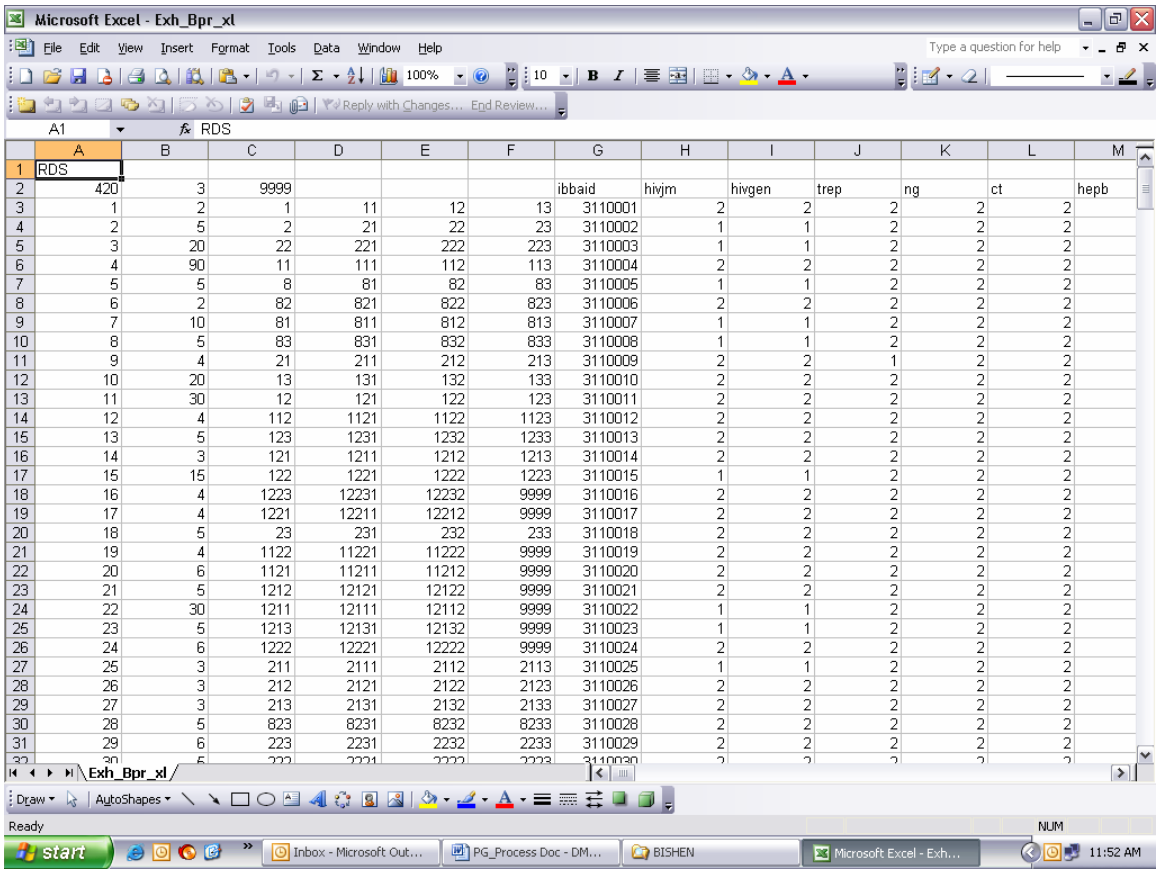
3.7 Data Analysis

After all the processes of entering, comparing, cleaning, merging, weighting of data were done for each district specific group data three levels of data analysis took place

- Top line analysis of data was done to show the finding for most important indicators.
- Descriptive analysis was done based on a detailed analysis plan which included almost all the variables and they were mostly frequency distribution tables.
- Exhaustive analysis was done based on another set of analysis plan which dealt with cross tables with two or more variable with some statistical measures like mean, median, chi-square tests, confidence interval estimation etc..

Analysis was done in SPSS for non-RDS groups by applying appropriate weights and developing syntaxes for generating tables. RDSAT was used for RDS groups and tables were developed based on the weighted estimates from the analysis. The top line analysis plan was prepared considering most important indicators e.g profile of the respondents, biological parameters, sexual behaviors related to consistent and last time condom use etc.

For data analysis, SPSS software could be run directly on the clean databases. Of course, labeling of variables, labeling of the responses, decoding, recoding and creation of new variables based on the analysis plan were done before generating the tables. Analysis in RDSAT required more preparatory work before data could be analyzed. The data was to be converted into text form before RDSAT could be used to generate estimates and that also in a definite format. It is advisable to recode the variables of interest in SPSS into new variables before converting the data into text form to be used in RDSAT because there is no provision of generating new variables in RDSAT. Another step before converting into text file was that the SPSS file had to be brought into Excel and prepare the definite format for RDSAT. Below is the example of the format of the excel file.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	RDS												
2	420	3	9999				ibbaid	hivjm	hivgen	trep	ng	ct	hepb
3	1	2	1	11	12	13	3110001	2	2	2	2	2	
4	2	5	2	21	22	23	3110002	1	1	2	2	2	
5	3	20	22	221	222	223	3110003	1	1	2	2	2	
6	4	90	11	111	112	113	3110004	2	2	2	2	2	
7	5	5	8	81	82	83	3110005	1	1	2	2	2	
8	6	2	82	821	822	823	3110006	2	2	2	2	2	
9	7	10	81	811	812	813	3110007	1	1	2	2	2	
10	8	5	83	831	832	833	3110008	1	1	2	2	2	
11	9	4	21	211	212	213	3110009	2	2	1	2	2	
12	10	20	13	131	132	133	3110010	2	2	2	2	2	
13	11	30	12	121	122	123	3110011	2	2	2	2	2	
14	12	4	112	1121	1122	1123	3110012	2	2	2	2	2	
15	13	5	123	1231	1232	1233	3110013	2	2	2	2	2	
16	14	3	121	1211	1212	1213	3110014	2	2	2	2	2	
17	15	15	122	1221	1222	1223	3110015	1	1	2	2	2	
18	16	4	1223	12231	12232	9999	3110016	2	2	2	2	2	
19	17	4	1221	12211	12212	9999	3110017	2	2	2	2	2	
20	18	5	23	231	232	233	3110018	2	2	2	2	2	
21	19	4	1122	11221	11222	9999	3110019	2	2	2	2	2	
22	20	6	1121	11211	11212	9999	3110020	2	2	2	2	2	
23	21	5	1212	12121	12122	9999	3110021	2	2	2	2	2	
24	22	30	1211	12111	12112	9999	3110022	1	1	2	2	2	
25	23	5	1213	12131	12132	9999	3110023	1	1	2	2	2	
26	24	6	1222	12221	12222	9999	3110024	2	2	2	2	2	
27	25	3	211	2111	2112	2113	3110025	1	1	2	2	2	
28	26	3	212	2121	2122	2123	3110026	2	2	2	2	2	
29	27	3	213	2131	2132	2133	3110027	2	2	2	2	2	
30	28	5	823	8231	8232	8233	3110028	2	2	2	2	2	
31	29	6	223	2231	2232	2233	3110029	2	2	2	2	2	

The first cell of the topmost row (A1) should have the word ‘RDS’. In the second row first cell (A2) should indicate the number of respondents, the second cell (B2) indicates maximum number of coupons distributed and the third cell (B3) is the code for missing responses due to skips in the questionnaire or any other reasons. None of the cells in the entire file can have missing cell. The respondents’ data starts from the third row. Starting from the third row, the first column is the default serial number of the respondent; the second column is the network size of the respondent (pulled from the RDS section in the questionnaire). The third, fourth, fifth and the sixth columns are the primary coupon number and the secondary coupon number for the corresponding respondents. Rests of the columns are the responses as per the questionnaire or recoded/created variables. Then

this is converted into a text file. It is to be noted that any deviations from this fixed format will make RDSAT unusable for analysis. Another limitation with RDSAT is that it does not generated tables in a printable format. So, manual imputation from the estimates generated in RDSAT is required to generate tables. Further, syntax cannot be generated in RDSAT and so for replication of the same analysis, one has to run the program again and again from the RDSAT menu.

In the next phase almost all the variables were taken into account and descriptive tables which were basically univariate analysis of the variables, were generated by developing SPSS syntax. The syntax with modifications was used for generating tables for other districts for the same group. Debugging of syntax was the iterative process after reviewing the tables generated.

Exhaustive tables were generated in the same process based on a separate analysis plan which was basically bivariate analysis with statistical tests like chi-square test confidence interval for the estimates. Complex sample module in SPSS was used to calculate these statistical estimates as cluster sampling was adopted in the survey.

There were several analysis plans depending on the requirement at various stages and data was analyzed at different times in different ways to fulfil the purpose. All these were done in SPSS for Conventional and TLC groups and RDSAT was used for RDS groups. Of course RDSAT software has limitations and does not have provision for generating bivariate tables with row and column percentages, statistical measures like weighted mean, median, chi-square test etc. but has other measures specific to RDS methodology.