An overview of Data Processing System of Survey data
(Indian Experience)

The System Design for data processing.

System Design of data processing is a scheme of actions to clean and tabulate the data collected from the respondants in a sample survey. It comprises of steps which are required to be followed in order to convert the collected data into meaningful information. It starts from the point of receiving the data at the Data Processing Center, and ends at dissemination of the unit level data and the generated tables to the users.

The System Design of data processing is different when the data is collected though paper schedule, than when data is collected using portable computers. In case of paper schedules, a few additional steps are required at the beginning to correctly transcribe the data into electronic media. Here we shall explain this kind of data processing.

In case of large-scale data processing, the total data processing work is structured over a number of functional steps. Data processing may decentralized over a number of Data Processing Centers (DPCs) and a number of people of different levels may be engaged in the data processing. Therefore, each step of system design should be planned beforehand, documented in details, concepts and definitions are explained, training workshops should be organised both centrally and then locally, and mid-course discussions should also be held to sort out unforeseen data problems. Hence, it should be a formal system of data processing.

The stages of data processing are :

i) Monitoring receipt position and Checking of Identification particulars.

(a) **Monitoring receipt position :** As the filled-in schedules flow in into the Data Processing Centers (DPCs), there should be a unit in the DPC to manage the receipt of incoming schedules. This unit is called Document Control Section (DCS). At the DCS, 'Receipt-Registers' are opened to record details of receipt position of schedules. The register is prepared on the basis of the information given in the sample list. As soon as the schedules are received in a D.P.Center, the number of filled-in schedules are counted and compared with those particulars given in the challan enclosed with the schedules and then these counts are to be entered in the 'Receipt Register' . In case of discrepancy, Field Offices are contacted.

In the DCS, a ‘Fly-Sheet’ is attached to one bunch of schedules (the FSU-level schedule and all the SSU-level schedules under it, comprise one bunch), after it is registered. Dates of completion of various stages of processing along with the related names of officials are recorded on the fly-sheet.

This Courseware has been received with thanks from the National Academy of Statistical Administration (NASA), CSO, India for use in the SAARCSTAT Web portal. Necessary editing has been done, wherever needed, without distortion of facts and figures.

(b) **Checking of Id-particulars :** This will inter-alia involve checking of Sample, Sector, State-region, district code, stratum-no, sub-stratum-no, sub-sample, Sub-round, Visit number, Village/block srl. no. and village size (frame population) from the sample list.

ii) Pre-Data-Entry (PDE) scrutiny of schedules for manual checking of obvious errors.

This checking is done to facilitate smooth data entry. Corrections are done on the schedules manually. If any errors of serious nature are detected, the same is taken up with field offices without delay. Number of PDE check points for a schedule may be around 50 to 60, which are sometimes prioritized for effective action.

iii) Data entry and 100% verification

A user-friendly SW should be developed using for this purpose. The SW should have provisions for screen-based Data entry, Verification and Updation job. The input screens are similar to those of the filled-in schedule in order and appearance. The SW should offer good navigation through the input screens and also through the data fields.

Verification means 100% re-entry of data and matching the second entry with the first entry for each of the data fields. This ensures that data has been accurately transcribed from the filled-in schedules into the computer. The is done with the help of the Data Entry SW. The SW ensures that unless data is fully verified, it won't be accepted as the valid input for the next stage of data processing.

Moreover, the Data Entry SW also should ensure completeness of the data entered in respect of the count of schedules as should be present in an FSU (Coverage Check). The count of second-stage schedules in an FSU should be automatically displayed on the first screen.

iv) Data structure and data consolidation

(a) The entered data may be re-formatted into text data by a data conversion SW. For text data, a detailed data layout should be prepared, which clearly shows the identification part and the data part of each data record.

Alternatively, entered data may be directly sent to a central database through LAN or website. Normally, data for one First Stage Unit (FSU) are sent at one time.

Data processing is usually more difficult if data is converted into text format. Therefore it is advisable to use a Relational DataBase Management System (RDBMS) for efficient storage and retrieval of data during its processing.

v) The Validation of data is done in three phases, namely, Content check (phase-I) Coverage check (phase-II), Howler check (phase-III).

Phase-I validation (Content Check) : This means running a series of validation checks on the data with the help of validation SW. These checks are also called Computer Scrutiny Program (CSP). The validation SW generates an error list showing errors or doubtful figures present in the data. The errors/ doubtful figures are checked from the filled-in schedules. If mistakes are found, corrections in the data are made. Phase-I validation run is executed three times, progressively rectifying the errors in each stage. This is required because manual rectification of errors leave scope for omission and even fresh induction of errors. Validation checks may be of several types, e.g. :

(a) fields are checked against list of admissible values (codes).
(b) arithmetic consistency checks, including range checks and subtotal checks.
(c) conditional checks involving different numeric/codified fields in a single or a group of records.
(d) matching of personwise or itemwise information furnished in different blocks.
(e) Checking of duplicate records

In some system, the Data Entry Operators(DEO) themselves carry out the validation checks and data revision. In other systems, the printed error-list is sent to Data Scrutinisers, who advise correction or confirm the figure, after consulting the filled-in schedule. In case, the filled-in schedule also fails to provide enough supporting evidences for necessary correction, the case is referred to the Data collectors in the Field Offices for a feedback.

Phase-II validation (Coverage Check) :

A master file is prepared containing all the FSUs surveyed, and having data fields to store counts of SSUs within each FSU. This special data file is called Directory. The entered data are checked against this Directory to ensure full coverage of all FSUs and all SSUs within each FSU. This check is called coverage check. Such checks include checking for

(a) full coverage of data vis-a-vis the Directory, in respect of each FSU and SSU.
(b) absence of any essential block of data.
(c) duplication of FSU or SSU-level data records.

Phase -III validation (Extreme value checking) :

Sometimes, even after validation checks, a few very extreme values (Outliers or Howlers) are left out in the data, which distorts the estimates. Detection such extreme values in data are called Howler check. This can be achieved by

(a) Checking the data field with reaspect to a suitable upper and lower bound.
(b) Sorting all the values of a data field in ascending or descending order, and taking 2 percentile and 98 percentile points as cut-offs.

(c) Computing some meaningful ratios (e.g. unit price of an item of consumption in a Sector x State, GVA per worker, Cereal consumption per capita etc.) and sorting the data file on those ratios to detect outliers.

vi) Computer-editing or Auto-correction :

After validation, necessary changes in the data are made with the help of a software to make it internally consistent. A set of rules are prepared for this purpose. These are called Computer Edit Rules. Here, changes in the data are made without referring to the filled-in schedules. As a matter of principle, such corrections are resorted to, to the minimum extent possible.

(a) all invavlid codes are replaced by blank.
(b) all subtotals/totals are computed for each data record, wherever required.
(c) additional records are generated if necessary.
(d) imputation is carried out to fill in blank field (e.g. value is imputed from quantity for each item of consumtion in Household Consumption Expenditure schedule)

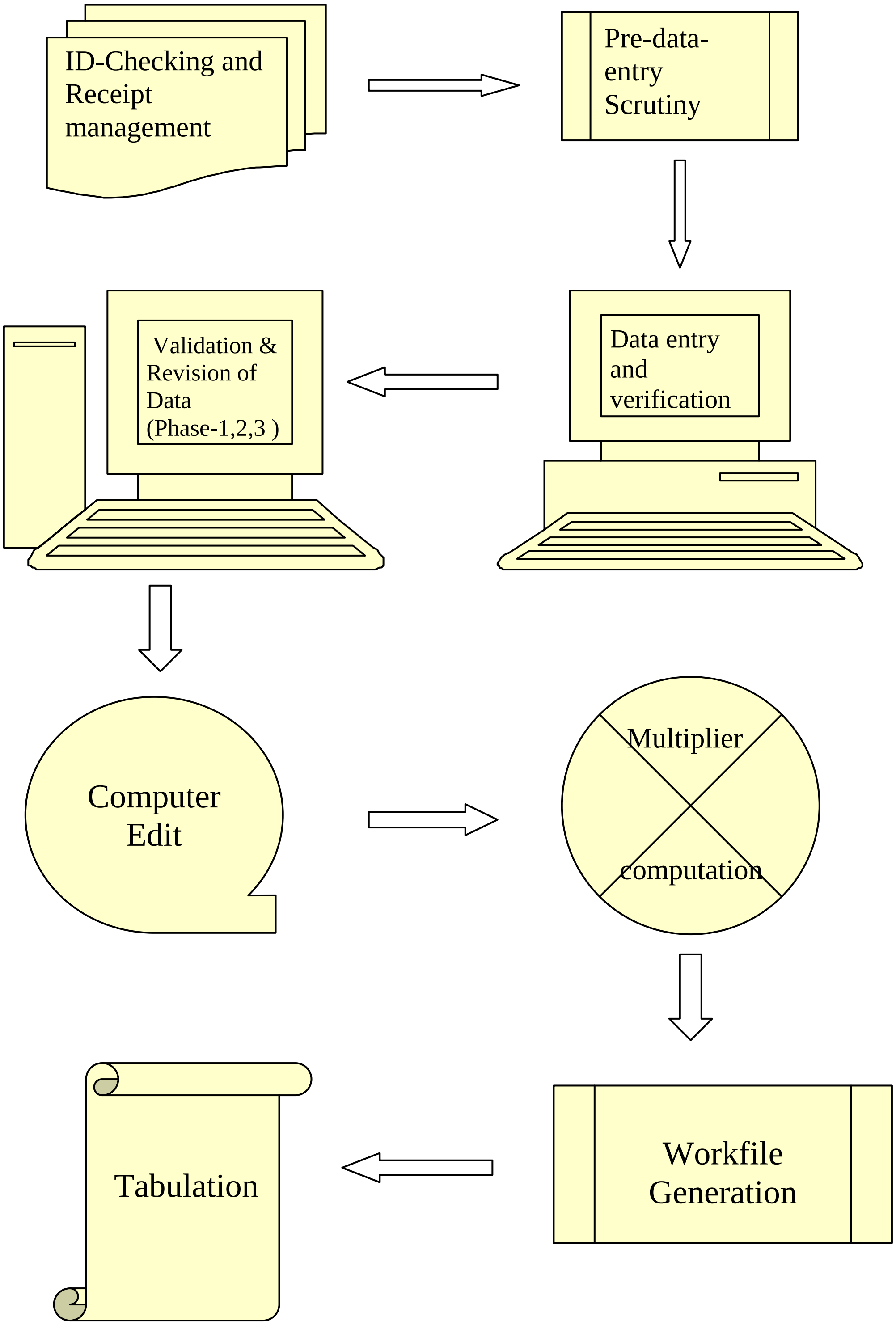
Auto-correction may entail creation of new records or deletion of existing records to restore inter-block consistency. Sometimes, in extreme cases, all the records of an SSU are deleted.

vii) Preparation of multiplier files, i.e. calculation of weighting factors for each Ultimate stage Units as per sample design. Computed multipliers should be independently checked for its correctness, before they are posted in the data for estimation and tabulation.

viii) Preparation of work files : After computer edit, unit level data is ready. But generation of tables directly from the unit level data is not very convenient. Therefore, suitable extracts from the unit level data are taken out to facilitate table generation. Related tables are usually grouped together, and all the data fields required to generate those related tables are extracted into a single file or table or view. These are called workfile. Typically, one workfile may be related to one group of entities, e.g. Person-level workfile, Household-level workfile and so on. Text workfiles are uni-format, and therefore, can be easily uploaded to any standard tabulation SW.

ix) Tabulation of data : Commercial Tabulation SW may be employed to generate tables from workfile. A table is nothing but a furnished statistical statement showing estimates of a variable distributed over different classes or groups or geographical areas, e.g. District-wise estimates of literate persons in the country.

Flow Chart of System Design for Data Processing


ID-Checking and Receipt management
Pre-data-entry Scrutiny
Data entry and verification
Validation & Revision of Data (Phase-1,2,3)
Computer Edit
Multiplier
computation
Workfile Generation
Tabulation