NATIONAL HOUSEHOLD SURVEY CAPABILITY PROGRAMME

Survey Data Processing: A Review of Issues and Procedures

UNITED NATIONS

DP/UN/INT-81-041/1

NATIONAL HOUSEHOLD SURVEY CAPABILITY PROGRAMME

Survey Data Processing: A Review of Issues and Procedures

UNITED NATIONS
DEPARTMENT OF TECHNICAL
CO-OPERATION FOR DEVELOPMENT
and
STATISTICAL OFFICE

New York, 1982

PREFACE

This is one of a series of studies designed to assist countries in planning and implementing household surveys in the context of the National Household Survey Capability Programme. The United Nations revised Handbook of Household Surveys is the basic document in the series. It provides technical information and guidance at a relatively general level to national statistical organizations charged with conducting household survey programmes. In addition to the Handbook, a number of studies are being undertaken to provide reviews of issues and procedures in specific areas of household survey methodology and operations. The major emphasis of this series is that of continuing programmes of household surveys.

The content, design and arrangements of household survey programmes will differ from country to country, reflecting varying needs, circumstances, experiences and availability of resources. In studying the options available to them, countries will choose those which make the best use of their resources to meet their specific needs. The objective of these studies is to examine the factors involved in these choices, and to promote good practices in the design and implementation of household survey operations. The present study deals with important organizational and technical considerations in the design and implementation of data processing procedures for household survey programmes. It provides an overview of the recent trends in data processing technology, and places this in the context of the existing situation in national statistical offices in developing countries. An attempt is made to identify the considerations involved in the choice of appropriate strategies so as to ensure timely processing of survey data, and at the same time to enhance national capability in the area of data processing. A review of the commonly used software packages is also provided.

Apart from the Handbook of Household Surveys, the reader may find it useful to refer to other related United Nations publications dealing with statistical organization or methods, such as the Handbook on Statistical Organization.* Data processing issues are specifically dealt with in Chapter 10 of this publication ("Computer Organization and Computerized Data Systems") and a number of closely related organizational and management issues are dealt with in other chapters. Where the data processing equipment to be used in the survey programme has been or will be acquired in connection with the population census, the relevant sections of Principles and Recommendations for Population and Housing Censuses** (particularly paras. 1.107-1.110 and paras. 1.131-1.142) should be consulted.

---

* Studies in methods, Series F, No. 28 ST/ESA/STAT/SER.F/28.

**Statistical Papers, Series M, No. 67 ST/ESA/STAT/SER.M/28.

In the preparation of this document, the United Nations was assisted by the United States Bureau of the Census serving as a subcontractor to the United Nations Department of Technical Co-operation for Development. The document was initially drafted by Ms. Barbara Diskin, with technical inputs from Mr. Robert Bair and overall direction of Mr. Robert Bartram. Subsequently, it was reviewed at a technical meeting on the National Household Survey Capability Programme held in April 1981 at New York, and revised at the United Nations Statistical Office in consultation with the United States Bureau of the Census. The document is issued in draft form to obtain comments and feedback, from as many readers as possible, prior to its publication in final form.

TABLE OF CONTENTS

*Sections of particular relevance to senior managers of the survey programme.

Page

Page

Page

I. INTRODUCTION

A. Objectives, Audience and Scope

1. Objectives and context

This document is one of a series designed to provide technical support to countries participating in the United Nations National Household Survey Capability Programme (NHSCP). Its objective is to provide an overview of the data preparation and processing task in the context of continuing collection of statistical data. More specifically, it addresses the various technical and operational problems of organization and implementation of data processing activities for household survey programmes undertaken by national statistical agencies in developing countries.

The context of this study is defined by the context and scope of the NHSCP. The NHSCP is a major technical co-operation effort in statistical development of the entire United Nations family. The main features of the Programme are:

(a) Country orientation, meaning that each project is designed to meet the actual data needs of a country in full consultation with national users, and also that full account is taken of differences in the existing levels of statistical development between countries.

(b) Leaving behind a self-sustaining survey-taking machinery, capable of providing a continuing flow of integrated data and also of meeting flexibly new data needs as they arise.

(c) Integration and co-ordination of statistical activity, both organizational and substantive.

(d) In the process of statistical development, focus on data collection from the household sector, in view of the key role of this sector in socio-economic activities of the population, particularly in developing countries.

Being a country-oriented programme, the NHSCP does not propogate any fixed model of surveys. The substantive content, complexity and design of the household survey programmes will differ from country to country, reflecting varying circumstances, experiences and availability of resources.

The general approach of the NHSCP applies also to the field of data processing. Consequently, the orientation of this study is that in developing and implementing a successful data processing system, each country must take stock of its own needs and potentialities. In studying the alternatives available to them, countries have to seek those which make the best use of their resources and meet their specific needs, and at the same time contribute to the development of enduring national capability in the field of data processing.

The study, therefore, discusses the various factors involved in making appropriate choices, rather than recommending any single approach. At the same time, its objective is to promote good practices in the design and implementation of procedures for statistical data processing.

2. Audience

This document is intended for use by designers and managers of household survey programmes, including subject-matter specialists, survey statisticians as well as data processing experts themselves. In the definition and detailed specification of the data processing task, it is essential to ensure close collaboration between data processors and non-data processors, and to develop a common understanding of the problems of taking raw survey data through the process of scrutiny and correction up to the analysis and reporting stages. The subject-matter specialists are concerned with the objectives, procedures and requirements of data collection, analysis, dissemination and use. They specify what needs to be done at the processing stage to meet these objectives. In this, they need to consult the data processing specialists, who are concerned with what can be achieved and how that is to be accomplisehd, taking into account the capabilities and limitations of the processing environment. The combined effort and knowledge of both the disciplines is necessary to develop the appropriate strategy, procedures and detailed specifications for the task. For this they must establish means of communication understandable to both, and develop proper appreciation of each other's task.

The managers of the survey programme are concerned with the overall design, planning and control of the whole operation, including data processing activities. Sections of this document of special relevance to senior managers are marked with an asterisk (*) in the contents.

3. Scope

As noted above, the specific context of this study is data processing for continuing programmes of household surveys, undertaken to meet national data needs as well as to enhance survey

capability. At the same time, many of the issues addressed here are common to any large-scale statistical data collection operation of the type usually undertaken by national statistical organizations. What is of fundamental importance is the context within which the data processing task has to be performed. National statistical organizations in developing countries typically face a number of problems which tend to result in data processing becoming one of the most serious bottle-necks in the entire process of statistical data gathering. Inadequate computer hardware and software, insufficient budget, scarcity of trained staff and difficulties in retaining experienced personnel are the most common problems. Difficulties are also caused by the lack of proper planning and operational control, overambitious work programmes and gross underestimation of the time required to do the job, and generally, by lack of balance between the rate at which data are collected and the rate at which they can be processed. Often insufficient attention is given to the design of survey questionnaires and clarity of editing rules and tabulation plans, reflecting lack of familiarity of subject-matter specialists with computer operations and lack of involvement of data processing experts in survey design. In many circumstances, problems also result from excessive dependence on external expertise and inputs, and failure to adopt the most appropriate strategy in the given circumstances. These problems are by no means all of a purely technical nature. Proper planning, organization and operation control - and, of course, good communication between data processors and other survey experts - are often the most crucial elements. A discussion of these problems and possible strategies to overcome them is the central theme of this document.

Apart from the problems common to all statistical data processing, household surveys have their own specific requirements. Household surveys are conducted to obtain a variety of data from the general population by enumerating a sample of households. Typically household surveys involve personal interviewing of respondents; the information collected may pertain to individual persons, to households or to aggregates of households or to any combination of these. Compared with large-scale censuses at least, sample sizes for household surveys tend to be relatively small, but the questionnaires used tend to be more complex and elaborate. These differences can have important consequences for the choice of the most appropriate procedures for household survey data processing.

B. Outline of the Contents

The present study begins with a description of the various steps in household survey data processing (Chapter II); basic requirements for assuring the processability of the questionnaires and technical considerations such as questionnaire design and layout, procedures for coding, editing and imputation, tabulation, and microdata linkage are discussed.

Chapter III considers organizational considerations and the operational and quality control measures required for successful

implementation of the data processing phase. Many statistical organizations can substantially increase their data processing efficiency simply by ensuring better utilization of existing facilities through proper planning and administration. Adequate attention needs to be paid to the organization and operational aspects. One of the areas most critical to the success especially of a continuing survey programme is the ability to stay within the projected budget. This will require a careful study of the costs involved. The budget should reflect the detailed data processing plan and should be based on carefully established estimates of workloads, rates of production, personnel available and costs of training, equipment, facilities, and supplies. Developing a realistic schedule is as difficult and important as arriving at a realistic budget and cannot be dictated by wishful thinking. In fact, the preparation of a calendar of activities and budget are interrelated and in both cases the first step is to spell out in detail the activities entailed in processing the data. While some activities such as manual editing and coding can be accelerated by an increase in the number of clerical staff, other data processing tasks are constrained by the availability of equipment and professional staff. Further, the various activities have to be performed in a logical order, though overlapping of various phases is possible and often desirable.

Chapter III also discusses the equally important control measures to be applied at the implementation stage. Procedures are needed for quality and operational control of manual and machine operations as well as of computer hardware and software. Maintenance of system, operational, and procedural documentation is one of the key factors in successful data processing for continuing survey programmes.

Chapter IV and V provide an overview of the recent trends in data processing technology, and place these in the context of the existing situation in national statistical offices in developing countries. It is useful to understand current trends in areas such as data entry, hardware and software in order to assess the level of data processing technology in a particular country and to consider areas of future expansion in order to strengthen data processing capability. Although many state-of-the-art techniques are discussed in this study, it is realized that the establishment of a country capability which can continue to function independently overrides by far the importance of utilizing the most modern techniques.

With this background, Chapter VI discusses the considerations involved in the choice of an appropriate strategy so as to ensure timely processing of the data generated by continuing survey activity, and at the same time to create or enhance data processing capability. The building of capability requires the acquisition and proper organization of data processing facilities, choice and development of appropriate software, and recruitment and training of good quality staff. A particularly important issue discussed relates to the provision of software: namely, the question of in-house development of custom software versus the acquisition of

purpose-built packaged software. Most statistical offices in developing countries do not possess resources to contemplate large-scale in-house development of custom software, and the appropriate strategy for them is to acquire existing software packages where available. This, however, does not imply that there can be no problems in the choice and operation of appropriate packages, or that suitable packages are always available to meet all data processing needs for continuing household survey programmes.

The discussion in Chapter VI is supplemented by a fairly extensive review of the available software packages in Annex I. The specific packages reviewed are selected on the basis of extensiveness of their use in statistical offices, their portability to different computer configurations and suitability for performing the required tasks in the circumstances and environment typically encountered in national statistical offices.

C. Some Factors Affecting Magnitude of the Data Processing Task

In addition to the obvious factors such as length and complexity of the survey questionnaires, sample size and design, survey timing and scheduling, the particular arrangements in a continuing programme can profoundly affect the magnitude and complexity of the data processing task. While survey design and arrangements are determined by numerous practical and substantive considerations in the light of users' requirements, their implications for the data processing phase need also to be kept in mind.

As noted earlier, the NHSCP, being a country-oriented programme, does not propagate any fixed model of surveys. The substantive content, complexity and design of the household survey programmes will differ from country to country, reflecting varying circumstances, experiences and availability of resources. A few examples of the survey major design factors affecting data processing are highlighted in the following paragraphs.

The size and complexity of the questionnaire have a marked effect on every aspect of the data processing system. The presence of many open ended questions increases the time and effort required during the coding operation, and the programs to edit and tabulate the data become more difficult to write and test as the questionnaire increases in size and complexity.

The sample size and the specific sampling design employed also affect data processing. For example, some software packages are not suitable for complex designs. Also, when units have been selected into the sample with non-uniform probabilities, the

resulting data have to be appropriately weighted before tabulation and statistical estimation. This generally increases complexity of data processing.

In data processing for continuing and integrated household survey programmes, a number of additional considerations are involved. "Continuing surveys" have been described as follows (United Nations, 1964, p. 3). "The most usual example of these surveys is where a permanent sampling staff conducts a series of repetitive surveys which frequently include questions on the same topics in order to provide continuous series deemed of special importance to a country. Questions on the continued topics can frequently be supplemented by questions on other topics, depending upon the needs of the country." This model represents a common - though not a universal - arrangement in NHSCP projects: the survey programme is often divided into (usually yearly) rounds, and a more or less substantial "core" of items is repeated in each round with varying "modules" added from round to round. Insofar as a significant core of questions can be kept unchanged in content as well as in layout from round to round, the data processing task is considerably facilitated. However, especially at the initial stages of the programme, strong substantive reasons may exist to introduce changes (improvements), which can substantially increase the magnitude and complexity of the data processing operation. Of all survey design considerations, the periodicity of the survey round has perhaps the greatest effect on the data processing system because it sets bounds on the length of time in which processing of each round must be completed to avoid a backlog of unprocessed questionnaires.

In an ongoing survey programme, a variety of sampling arrangements are possible. At the one extreme, each survey or round may be based on an entirely different sample of households; at the other extreme, the surveys may be completely "integrated" in that "data on several subjects are collected on the same set of sampling units for studying the relationship among items belonging to different subject fields" (United Nations, 1964, p. 3). More commonly, different surveys or rounds may be integrated to varying degrees, for example, they may employ common sampling areas but different sets of sample households, or the sample of households may be partially, but not completely, rotated from one round to the next. The possibilities and requirements of data linkage and combined analysis across survey rounds will differ depending upon the particular sampling arrangement, as would the resulting data processing requirements and complexity.

Any specific survey or survey round may be more or less "multi-subject", that is "when in a single survey operation several subjects, not necessarily very closely related, are simultaneously investigated for the sake of economy and convenience ... the data on different subjects need not necessarily be obtained for the same set of sampling units, or even for the same type of units" (United Nations, 1964, p. 3). As a result, the survey may involve more than one type of questionnaire, considerably affecting the programming

and other data processing requirements. Furthermore, data from different types or levels of units (such as communities, households, holdings, individuals) may need to be pooled or linked together, which will tend to complicate the structure of the resulting data files. The system used to process the data must take into account the reporting and reference units in the survey in order to decide on data file structure and to assure that the proper codes are supplied to get from one file to another. In addition, the specific reporting and reference units used may affect the ability to link the data collected across survey rounds or to other sources of data. It should also be noted that many software packages cannot handle complicated "structured" files.

II. DESCRIPTION OF THE PROCESSING TASK

This chapter describes the various components or steps that comprise the data processing task for household survey programmes, from planning and questionnaire design to data linkage and system documentation. The objective here is to discuss some important technical considerations in the design of data processing procedures; organizational and operational considerations in the implementation of the task are taken up in the next chapter.

A. Planning for Data Processing within the Total Survey Programme

For an organization planning to undertake a continuing programme of statistical data collection, the objectives and scope of the programme must be determined, taking into consideration the capacity for timely processing and dissemination of the data in a realistic manner. Hence, a primary task of the data processing managers and specialists is to participate in the overall planning of the survey programme. On their part, those responsible for data processing must realize that data needs of the users should be the first and foremost consideration, and that data processing is a service provided to meet these needs as well as possible. On the other hand, the subject-matter specialists must ensure, and data processors must insist, that in deciding upon the data collection programme, the processing task must not be allowed to become unmanageable. The rate at which data are collected must be compatible with the rate at which they can be processed and utilized. The failure to produce results in a timely fashion reduces, or can even destroy, the worth of the data and is bad for the morale and reputation of the statistical agency. Processing data from household surveys demands considerable sophistication, and the resources and level of personnel potentially available determine to a large degree how ambitious the data processing, and hence data collection, plan can be.

Initial data processing plans should be drawn up when the surveys are first designed and the project time-table and budget calculated. At this time, the various tasks to be performed should be identified and flow charts of the data processing steps drawn.

The capacity of the existing facilities including hardware and personnel should be evaluated and plans for upgrading them formulated. Existing software should be evaluated in the context of the tasks to be performed, and decisions made on the extent to which available general purpose packages can be used, and the extent to which it would be necessary to modify the existing, or to develop new, special-purpose software. Very serious consideration needs to be given to the substantial time and resources which any new software development effort is likely to demand.

As noted above, one of the most serious problems to be avoided in continuing programmes of household surveys is the piling-up of unprocessed data. The survey time-table should take into full account the estimates in person days and also in elapsed time required for preparing documentation, writing and testing of programs, and performing the actual data processing. It is essential to make these estimates before the time-table, complexity and sample size for the surveys are fixed.

Detailed formulation of the plan and requirements of data processing is in fact a continuing operation which needs to be accomplished prior to implementing any specific step of the operation such as data entry, program development, editing and coding and tabulation for any particular survey. Another continuing requirement is that up-to-date and complete documentation of all procedures, operations and programs is maintained.

It should also be emphasized that in planning and working out of details of the data processing procedures, close co-ordination is essential among all persons - managers, subject-matter specialists and data processing experts - working on the survey programme. This co-ordination would include agreement on what outputs are necessary before computer programs are written. As the data processing system is being tested, the subject-matter specialist should be encouraged to review the output to check its accuracy and adequacy. As the survey data are being processed through the editing and tabulation phases, it is advisable to provide for numerous opportunities for verification to assure that the results meet the needs of the users and produce statistically correct results.

B. Assuring Processability of Survey Questionnaire

Once the content of individual surveys in the programme has been determined, the data processing experts should be fully involved in the design of survey questionnaires so as to ensure their processability. The guiding principle in the design of questionnaires should be to collect the most accurate data possible, but the convenience of data processing must also be given its due importance. Of course, in the case of serious conflict between data

collection and data processing requirements, priority should be given to the former; it should however be appreciated that there are forms of questionnaire design and coding schemes that lead to simplification of the data processing task without adversely affecting the field work (Rattenbury, 1980, p. 12). In fact a well designed questionnaire layout can assist both the operations. For example, the use of shading, or different colours if feasible, can assist the interviewer to distinguish the responses to be entered in the field from the items to be coded in the office. Close collaboration between survey designers and data processing personnel is, therefore, a clear necessity. Further, data processing personnel often possess special skills in designing neat forms and questionnaires, and use should be made of these skills where possible.

1. Identification of records

Careful consideration must be given to designing the system of identification numbers for survey questionnaires. This is particularly important for programmes of related surveys where data from a number of individual surveys as well as from other sources are to be linked. Each survey questionnaire must have a unique identification number which appears on a conspicuous location on the title page of the questionnaire. An inappropriate system of questionnaire identification can result in serious difficulties in performing operations such as sorting of data, microlevel linkage of data from different sources and estimation of sampling and response variances.

The system of identification must define all that is necessary to locate each survey questionnaire in the total data set. For example:

(a) The survey programme may consist of a number of "rounds", in which case an indication of the round number should appear as a part of questionnaire identification. Similarly, when survey rounds are divided into subrounds the latter also need to be identified.

(b) Sufficient information should be provided to identify the sample structure (such as the domain, stratum and cluster), as well as the administrative area if relevant, to which each enumerated unit in the sample belongs. It is only on the basis of such information that sampling variances can be computed. Where a series of surveys is based on the same common set of sample units, it should be possible to link the information for these units across different surveys.

(c) Any given survey may involve a hierarchy of questionnaires, pertaining to related units at different levels. For example, there may be a questionnaire for each sample area or

community, followed by household questionnaires, and within each household, questionnaires for each household member. In the above example, the identification number for individual members may consist of codes for the survey round, the sample area, the household and finally individual member within the household; identical round, area and household number would appear for the corresponding household so as to permit direct linkage of the household and individual member data. In fact, the identification numbers should be defined in a way to permit sorting and linkage of the entire data file for various survey rounds and levels of units in any required order, using common data fields in a fixed location for sorting and linkage.

(d) Frequently it is necessary to divide a questionnaire into "record types", such as 80-column cards or card images on disk or tape. It will then be necessary to include the record type as an element in the system of questionnaire identification. Also, a clear indication of the record type should be provided: for example, natural breaks in the questionnaire such as new sections or pages should preferably form the beginning of new record types.

Two points of practical significance may be noted in the choice of identification number. First, it is desirable to avoid the use of non-numeric characters. Secondly, it may not be possible to provide all the necessary information for record linkage as a part of questionnaire identification without making the identification number too long. For example, a census may use a complex system of uniquely identifying enumeration areas which specifies the various administrative and geographical units to which the area belongs; in a sample survey, by contrast, a much smaller number of area units may be involved and a simple sequence of numbers may suffice to identify sample areas uniquely. The linkage of census and survey data at the area level may then be achieved through a conversion table which maps the simpler survey area identification numbers onto the more complex system for the census. The detailed sample structure may be specified and mapped onto the simple area identification numbers in a similar way. At a later stage in the data processing operation, the more complex area identification numbers may be transferred onto individual questionnaires as new data fields. In a questionnaire divided into a number of record types, the identification number needs to be repeated on each record type.

2. Precoding and layout

Responses during an interview can be recorded in various ways such as checking a box or circling a code, writing in a code or a number, or writing in the information in words, either verbatim or in a condensed form. More specifically, five forms of recording may be distinguished (World Fertility Survey, 1976):

(a) Fixed-alternative questions: in this case all possible or alternative answers are predetermined (such as yes/no) and the interviewer simply checks or circles only one of those.

(b) Multi-coded questions: same as above, except that the interviewer checks or circles as many codes as apply. An example is questions enumerating reasons for something, when more than one reason may be given by the respondent to a particular question.

(c) Number or value questions: here the answer is specified as a numeric value which can be directly used as the code. Examples are age, number of persons.

(d) Open ended questions: here the response is descriptive either because the possible answers are too many to be precoded, or are too complex or unknown for this purpose.

(e) Semi-open ended questions: these represent a mixture of types (a) - (c) with type (d). Ideally, the "fixed" part covers the great majority of responses, but provision exists for recording open ended responses where necessary.

It is usually more efficient to adopt alternatives (a) - (c) since they take less space, require less time during the interview and reduce the amount of coding required in the office. By contrast, the coding of open ended questions can require substantial time and effort depending upon the complexity of the code involved. Sometimes it is not possible to avoid open ended questions without sacrificing the completeness or richness of the responses. However, an effort should be made to keep the number of open ended questions within limit, and to choose the "fixed" or at least the "semi-open ended" form where possible.

The physical layout of the questionnaire and coding scheme used affect both the speed and accuracy of the coding and data entry. Crowding of questions to save space at the expense of readability should be avoided. Check boxes and coding boxes should be in logical positions which relate in an obvious way to the corresponding question. When check boxes are provided, preprinted codes should appear alongside. Coding schemes should be consistent; e.g., if a "Yes" is coded as "1" and a "No" as "2" for one question, it should be coded in the same way for all such questions. Items to be coded in the office, as opposed to those answered in the field, should be clearly labelled to that effect. Sufficient space should be allowed for the codes supplied in the office and these codes should not in any way obscure the original entry. If possible, the format should allow the data entry operator to pick up the data in a

uniform pattern, either in a continuous flow from left to right (or right to left if language so dictates) or from top to bottom. There should be no need to flip back and forth within the questionnaire or to unfold pages. If shading or use of colour is a possibility, the data to be entered can be highlighted clearly to aid the data entry operator. In summary, questionnaire design is greatly affected by coding procedures and requirements and by data entry considerations. These operations are discussed in the following sections.

C. Coding

Coding is the process in which questionnaire entries are assigned numeric values. The objective is to prepare the data in a form suitable for entry into the computer. The coding operation may involve one of the three alternatives:

(a) Assigning numerical codes to responses recorded in words or in a form requiring modification before data entry. These include items such as geographic location, occupation, industry and other open ended questions. This is "coding" proper.

(b) Transcription, in which numeric codes already assigned and recorded during the interview are transferred (rewritten) on special spaces provided in the questionnaire or onto separate coding sheets. The objective is to facilitate data entry.

(c) In certain cases no coding or transcription is required, that is, the numeric responses recorded by the interviewer are directly used for data entry.

In general, transcription should be avoided whenever possible because it is time-consuming, and more importantly, introduces new errors into the data. Alternative (c) is of course the most economical and error free, and is frequently followed when the data have been recorded in a simple tabular form, as for example in household rosters enumerating basic population characteristics.

Whenever possible space for coding should be provided in the questionnaire itself rather than using separate coding sheets which are cumbersome and are easily lost. The first alternative is generally simpler and less prone to clerical errors, and should be followed if sufficient space is available to do this without adversely affecting the clarity and layout of the questionnaire.

It is generally preferable to code all the questions recorded in the questionnaire separately, and not to condense the range of

responses to a question at the coding stage. It may seem that some questions need not be coded at all, or that space can be saved by combining several questions into one code. However, such false economy can result in loss of valuable information, and certainly increase the risk of making errors. Furthermore, some redundancy in the coded information can be useful in checking internal consistency of the information recorded.

All questions which have answers in the same range (e.g. questions with yes-no responses) should have these coded in the same way and same order. Categories common to all questions, such as answer "not known" or "not stated" should be coded in a standard way.

A manual should be written to give explicit guidance to those persons involved in coding. It should give examples of each task to be performed and should leave no doubts in the mind of the person involved. The manual should be written by the supervisor responsible for the coding operation and verified by the survey manager and data analysts to ensure that the coding scheme is compatible with analytic requirements, and that it is consistent between surveys in the programme.

D. Data Entry

Data entry refers to the transference of data to a computer readable medium. Two approaches are possible:

(a) Operator controlled data entry, which involves keying-in of the coded data onto cards, tape or disk.

(b) Reading of the data directly by optically scanning questionnaires or coding sheets.

1. Operator controlled data entry

Operator controlled data entry is the more common and generally more appropriate approach in developing countries, especially for household surveys where the volume of the data involved at any given time is likely to be less than, for example, in the case of a full-scale census.

There are basically two approaches in the design of operator controlled data entry: the use of various "fixed format" record types and the use of "source codes". The former is the more traditional approach, whereby the questionnaire is broken down into a fixed number of record types and each item of data is located in a fixed position on a particular record type.

An illustration of fixed format is given on page 15. The illustration refers to the agricultural holdings of a household. Note that columns 1-13 are fixed locations for the identification of the questionnaire and that the record type is indicated on the upper right-hand corner by "Form" and "Card" with the values "1" and "1", respectively. Position numbers are associated with fields on the page, indicating its precise location on that particular record type. Thus, the reference person's surname will always appear in positions 46 through 65 on record type "1". Record type two, indicated by a "2" in position 15, begins after the name information. The remainder of the record identification will be automatically repeated in positions 1-14 of record type 2. Positions 22-27 indicate that the reference person has 200 cattle and this is keyed accordingly. Since the respondent has no sheep or lambs, positions 28-33 must be left blank by skipping over them to positions 34-39 where 15 goats or kids are recorded as this example. Thus, all positions must be accounted for, either by entering the information recorded on the questionnaire or by skipping the blank positions which do not apply. There may be additional record types which are totally blank because the items were not applicable; in such cases it is not necessary to key the records at all.

The use of source codes implies the assignment of a unique code to each item of data in the questionnaire. An illustration of the use of source codes is given on page 16. This is just one page of a lengthy questionnaire which bears an identification number on the front page. The source code appears to the left of each response as a three-digit encircled code. The data are entered as a series of source code and value fields, where each field is of equal length and only those items or source codes which actually have values are entered. Suppose that the illustrated questionnaire has an eight-digit identification code of "10634021" on the front page and that the maximum length of any response is six digits. Then, the data would be entered as the string of numbers 10634021-374-000010-375-001500-376-000015-377-001000-378-000003-380-000001-etc." (The dashes are shown only to separate fields and would not be keyed.) Note that "379" is not included because there is no associated value. This approach may generate multiple records. Each record will contain the unique questionnaire identification number in the first eight positions and have a format identical to all other records. One of the first steps in processing data entered in this way would be to reformat them, using the source codes as pointers into a vector or an array and storing the associated values in their respective locations.

There are advantages and disadvantages to either approach. If most of the questions are applicable to all respondents, the fixed format data entry approach can be more economical in terms of space required on the data entry medium since the data are packed together; i.e., one-digit responses require only one position, three-digit responses three positions, etc. In addition, there is no need to reformat the data before further processing. However, blank fields within record types must be keyed, and there is a much greater chance of erroneously shifting data during data entry.

Example of fixed format questionnaire

AGRICULTURAL UPDATE

IDENTIFICATION NUMBER

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
4	2	1	3	5	0	6	0	2	1	0	3	6	1	1

A. Village name

OFFICE USE: 16 17 18 19 = 0 2 1 4

B. Enumerator's name

1a. Does anyone in this household normally do any farming, that is, grow rice, coconuts, or any OTHER crop or raise livestock?

20 — 1 [X] YES → *SKIP to 2a* 2 [ ] NO →

1b. Does anyone in this household raise any fowls, chickens, ducks, geese, turkeys, or guinea birds of all ages?

21 — 1 [ ] YES → **1c.** How many altogether? 22-27 [ ] Total — *If more than 99, continue with question 2a. If 99 or less, END INTERVIEW*

2 [ ] NO — *END INTERVIEW*

2a. Who is responsible for making day to day decisions for this farm?

Given name 28-45: R O B E R T

Surname 46-65: B R O W N

Call name 64-80: B O B

REFERENCE PERSON *for interviewing purposes*

15 — 2 **FIRST CHECK ITEM**

Is this person a household member? Individual No. Address

16 — 1 [X] YES — *Give individual number and address* → 17-18: 0 1

2 [ ] NO — *Probe to determine if any household member also makes the day to day decisions for this or any other farm. If there is, enter name in question 2a. If not — END INTERVIEW.*

2b. Does anyone in this household normally do any OTHER farming SEPARATE from *(the Reference Person's)* operation?

19 — 1 [ ] YES → **2c.** Who is responsible for the day to day decisions for each additional farm? *(Write the name of the operator for each farm on a separate continuation sheet (Form CS-2, item 2a), then continue with question 3a for the Reference Person in 2a above.)*

2 [X] NO — *SKIP to 3a*

3a. Now I have a few questions about *(the Reference Person's)* operation.

Is there anyone OUTSIDE this household who is also responsible for the day-to-day decisions for this farm?

20 — 1 [ ] YES → **3b.** What are the names and addresses of all these decision makers?

2 [X] NO — *SKIP to 4*

Given name	Formal name	Call name	Address

3c. Of all of these persons, INCLUDING *(the Reference Person)*, who is the oldest?

Name of oldest person

SECOND CHECK ITEM

The oldest person is —

21 — 1 [ ] The reference person in 2a — *Continue interview*

2 [ ] Outside the household — *END INTERVIEW*

4. Is *(the Reference Person)* responsible for —

a. Any cattle of all types and ages? [X] YES → How many? 22-27: 0 0 0 2 0 0

b. Any sheep or lambs of all ages? [ ] YES [X] NO → 28-33

c. Any goats or kids of all ages? [X] YES → How many? 34-39: 0 0 0 0 1 5

d. Any pigs or piglets of all ages? [X] YES → How many? 40-45: 0 0 0 0 1 0

e. Any donkeys or horses of all ages? [X] YES → How many? 46-51: 0 0 0 0 0 8

f. Any fowls, chickens, ducks, geese, turkeys or guinea birds of all ages? [X] YES → How many? 52-57: 0 0 0 2 0 0

5a. Altogether how many acres of land does *(the Reference Person)* own, rent, lease from others, or otherwise use? DO NOT INCLUDE the house lot or land rented out or used by others.

58-64: 0 0 1 0 0 . 0 0 Total acres (Whole . Decimal)

5b. How many other acres does *(the Reference Person)* use which have not been mentioned as yet?

65-71: 0 0 0 1 0 . 0 0

Example of source coded questionnaire

Section VI – LIVESTOCK AND POULTRY (Day of Interview and 1978) (Continued)

CATTLE

71a. How many of your cattle or cows of all kinds or ages did you or other members of this household SELL in 1978?
(374) 10 Cattle
☐ None – SKIP to Q. 72a

b. How much money did you get for those sold?
(375) $ 1500 . 00 Total

72a. How many cows or cattle did you or other members of this household BUY in 1978?
(376) 15 Cattle
☐ None – SKIP to Q. 73

b. How much money was paid for them?
(377) $ 1000 . 00 Total

73. How many cows or cattle of all ages and kinds DIED?
(378) 3 Cattle

74. How many were stolen or lost?
(379) _____ Cattle

75a. Did you or any other member of the household milk any cows in 1978?
(380) 1 ☒ Yes
2 ☐ No – SKIP to Q. 78a

b. Were cows USUALLY milked every day of the year?
(381) 1 ☒ Yes – SKIP to Q. 76a
2 ☐ No

c. About how many months were cows milked in 1978?
(382) _____ Months

76a. How many cows were usually milked each day?
(383) 4 Cows
(to)
(384) _____ Cows

b. How much milk do you usually get per day when you are milking? Include both morning and afternoon milkings.
(385) 16 Pints
(to)
(386) _____ Pints

c. How many pints of milk FROM YOUR OWN COWS are usually used in the home per day?
(387) 16 Pints

77a. Was any milk from your cows sold in 1978?
(388) 1 ☐ Yes
2 ☒ No – SKIP to Q. 78a

b. How much was sold?
(389) _____ Pints
OR
(390) _____ (to) (391) _____ Pints per day

c. How much money did you get for this milk?
(392) $ _____ . 00 Total
OR
(393) _____ Cents per pint average

FORM APPROVED 111-S-78)

Furthermore, the source code approach, if applied to a questionnaire which has many blank items, can be more efficient in terms of space needed and is far easier to enter from the standpoint of the data entry operator.

Operator controlled equipment for data entry encompasses conventional card punch machines, variations of the card punch machine, and key-to-disk, -diskette, -cassette, and -tape machines. These machines vary widely in the length of the record they can produce, their ability to handle multiple formats and their editing capabilities. When designing the questionnaire, it is important to know which machine will be used for data entry, so as to take full advantage of its capabilities and, at the same time, not introduce a situation which is impossible to handle on the available machine.

If more sophisticated programmable data entry machines are being used, some data editing may be done at the data entry stage. It may be particularly appropriate to undertake a "format edit" (see next section) at this stage if appropriate facilities exist.

It is true that data from household surveys tend to be more complex but at the same time less voluminous compared for example to data from full-scale censuses. Nevertheless, data entry often proves to be a problem area because the magnitude of work is underestimated. Making a realistic estimate requires having the following information:

(a) Number of data entry stations available for this work.

(b) Number of shifts of data entry operators.

(c) Number of productive hours on each shift.

(d) Number of data entry operators on each shift.

(e) Average number of key strokes per hour.

(f) Number of questionnaires.

(g) Average number of strokes per questionnaire.

(h) Percentage of verification to be done.

Some assumptions must then be made, based on the local situation, about other factors which affect overall production, such as:

(a) Ten percent of the equipment may not be operational at any point in time because of mechanical breakdown or operator absence.

(b) Five percent of the data will have to be rekeyed because of errors encountered in verification.

(c) Keying of manual corrections during editing will be the equivalent of five percent of the original workload.

Suppose the following case exists:

(a) Ten data entry stations available for this work.

(b) Two shifts of data entry operators.

(c) Six productive hours per shift.

(d) Ten operators on each shift.

(e) Average of 8,000 strokes per hour.

(f) 10,000 questionnaires.

(g) 2,000 strokes per questionnaire.

(h) 100 percent verification.

Making the assumptions listed above for other factors affecting production, the calculation in terms of days is the following:

Number of work days=Total strokes/strokes per work day

=(No. questionnaires x strokes per questionnaire x verification factor x factor for rekeying for data entry errors x factor for keying for editing problems)

(No. stations x factor for stations operational efficiency x shifts per station x productive hours per shift x strokes per hour)

=(10,000 x 2,000 x 2 x 1.05 x 1.05)/
(10 x .9 x 2 x 6 x 8,000)

=(44,000,000 strokes)/(864,000 strokes per work day)

= 51 work days

This does not imply that 51 work days of data entry must precede the remainder of the processing. It says simply that the total processing cannot be accomplished in less than 51 work days given the information available and the assumptions made.

Once the quantity of data entry is established, a plan for accomplishing it can be set up. Explicit instructions must be written for the data entry operators and they must undergo training to assure that they understand their task. Quality and operational control measures (to be discussed in the next chapter) must be designed for the data entry operation and are central to its success.

2. Optical scanning

Optical scanning (or optical character recognition, OCR) is a more sophisticated technique for converting data to machine-readable form. Human-readable documents are optically scanned and read into the computer directly, without keying or rekeying. The data to be read must be placed in predetermined positions on the questionnaire.

The two common forms of the technique are the optical character reader (OCR) and the optical mark reader (OMR). The OCR reads hand-printed letters and numbers and converts them to codes. The OMR, which is more common of the two, translates responses marked on a piece of paper or card with a special pencil into specific numbers or letters. OMR systems can have certain advantages over other types of data entry, particularly where time and accuracy are important:

(a) Processing can be completed in less time since data entry is accomplished in one operation.

(b) The need for keying equipment and key operators is eliminated.

(c) More accurate data are produced since the information recorded is not subject to keying-in errors (United States Bureau of the Census, 1979, Part B, p. 123).

(d) Questionnaire identification numbers may be preprinted and read directly without the possibility of error.

(e) Versatility is another advantage of OMR's. Not only do all models produce one record per sheet read, but some can read multiple sheets, combine them in the proper order, and output a single record (Bessler, 1979, p. 74).

However, the use of optical readers is a sophisticated technique requiring precision in the design and printing of questionnaires, and the use of high quality paper and special ink for printing. This can substantially increase the cost of questionnaire printing. Two types of ink are necessary so that printed questions and answer boxes can be distinguished from the recorded responses and codes. The questionnaires need to be handled carefully in the field and office, which is not possible in many circumstances. Above all, the more conventional labour intensive methods of data entry may present no particular problems in many statistical offices in developing countries, while the more sophisticated approach may unnecessarily increase their dependence on imported goods and services.

Consequently many users continue to regard conventional keypunching as the most cost-effective method of data entry. However, with the improvement of OMR equipment over recent years, it may be becoming a more viable alternative for large-scale data entry.

E. Editing and Imputation

These processes are designed to check that the information contained in the questionnaire is complete, recorded in the prescribed manner and internally consistent; and to take appropriate action when these conditions are not fulfilled. Editing refers to checking and correction of data. Imputation is the process of filling in with plausible answers those data fields for which there are no responses, or substituting alternative answers for responses considered unacceptable on the basis of criteria of logic and internal consistency. It is clear that the conceptual and operational distinction between "editing" and "imputation" is not clear-cut.

It is important to the interpretation of the data that errors and inconsistencies are corrected before the analysis phase. The objectives of "cleaning" (editing, correction and imputation) of data are: (a) to enhance its quality, and (b) to facilitate the subsequent data processing tasks of recording, tabulation and analysis. It needs to be strongly emphasized that cleaning of household survey data is not a trivial task: it has frequently proved in practice to be the most time-consuming of all data processing tasks in survey. The role of this phase has to be viewed in relation to the time and effort required to accomplish it.

Undoubtedly the cleaning of data is an essential phase of the survey operation. However, the basic question to be considered is the extent or degree to which data cleaning should be carried out. On the one hand, all editing and imputation amounts to altering what was actually recorded in the field by the interviewer; inappropriate procedures can have serious consequences for the validity of the data. On the other hand, editing permits the improvement of clearly incorrect and incomplete data on the basis of their internal consistency and relationships. For these reasons it is necessary to discuss edit and imputation philosophy prior to offering suggestions for its implementation.

In household survey operations editing and correction generally need to be done at two stages: manually before coding and data entry, and subsequently by computer. Both stages are essential. Manual editing is required since, at the very least, questionnaires must be checked for completeness, legibility, identification and other important data items prior to coding and data entry. Computer, also called machine, editing is a more detailed and complete application of the same edit rules. It is preferable because of the possibilities of human error in, and inherent limitations of, the manual operation, and because of errors introduced during coding and data entry. An important consideration is the manner in which the task should be divided between manual and machine operations. Similar issues arise in relation to corrections following machine editing, which may be made manually or automatically.

1. Edit and imputation philosophy

The introduction of electronic data processing into statistical operations over the past three decades has vastly expanded the scope and complexity of editing and imputation. It has become a common practice in some countries to change the answers on each questionnaire that do not seem consistent with a respondent's other answers, and to fill in almost all omissions with substitutes that are deemed plausible (Banister, 1980, p. 1). The gradual expansion of computer editing and imputation procedures has caused some users and producers of data to be uneasy. After all, the purpose of collecting data is to discover or reaffirm some elements of truth about the population. To make extensive changes in the collected data prior to making them available for analysis is to violate a basic principle in data collection that the integrity of the data should be respected.

This section examines the arguments for and against elaborate editing and imputation.

Several arguments can be given in support of elaborate editing and imputation: these operations improve or at least retain data quality; make data more convenient for processing and analysis;

and enhance the credibility of the collected data in the eyes of the user. Complete imputation, i.e. substitution of all not stated values by imputed values, is justified on grounds of expediency. Some analysts feel that a column of "not stated" values in tabulations is not really informative, and that users in any case ignore these in interpretation of the data. Some analysts also feel that complete imputation is justified since often those who responded are similar to those who did not.

On the other hand elaborate editing and imputation can be criticized for several reasons: it can significantly change the collected data and introduce serious errors into the published data; it can destroy all evidence that particular data are of poor quality and should be used with caution or should not be used at all; it can suppress all anomalies in the data and fill in all unknowns, thus giving the user unwarranted confidence in poor data. There can be serious problems in particular with complete imputation:

(a) It may be impractical to use appropriate criteria to select reasonably informed substitute values for non-responses. Computer size, software capability, and programming complexity are all contributing factors.

(b) Users may not be given complete information, or may not choose to avail themselves of information concerning the degree and type of imputation that occurred in the processing.

(c) Good information could conceivably be destroyed by edit rules which were so strict as not to permit rare, but possible, cases, such as married 12 year-old female with a child.

(d) It is often difficult to know which of two inconsistent responses is wrong. It is possible to make an incorrect change resulting in consistent, but distorted, data. This problem is serious enough that a great deal of work has gone into devising ways to ensure that the incorrect response is the one that is changed (see, for example, Fellegi and Holt, 1976, pp. 17-35).

(e) The extensive use of editing and imputation can contribute to greater sloppiness in data collection. If enumerators know that all omissions and errors will be fixed by the computer, they may exert less effort at collecting high quality data.

In summary, while data cleaning is an essential step in survey data processing, it is necessary to be cautious against the overelaborate use of editing and imputation procedures. It is with this view that the possible advantages of data cleaning, noted earlier, will be examined below:

Improvement or maintenance of data quality. This depends to a large degree on the basis on which data are corrected or imputed. "Feedback editing", i.e. correction by going back to a previous stage of data collection or processing, can improve data quality. The form of feedback editing that can improve data quality most is field editing where anomalies and omissions discovered in the field are resolved by going back to the respondent for clarification. To a lesser degree, "consistency editing" can also contribute to data quality. This implies correction on the basis of logical and substantive criteria, internal consistency and other information available within the questionnaire. However, "blind imputation", particularly of entire questionnaires, can result in serious distortions in the data.

Data processing convenience. The presence of inconsistent and incomplete values in the data can substantially increase the complexity of the processing task. Program development, documentation, and tabulation layout all become more involved. The convenience of data processing can be a compelling argument in favour of complete imputation in situations where the incidence of inconsistency and incompleteness is sufficiently low so as not to affect the aggregate data significantly.

User convenience. It is true that in many types of data analyses, the presence of missing values can be a nuisance to the analyst. However, users as well as producers of the data should be aware of the fact that some data are better than others, that some questions are better than others, and that some questions simply elicit a higher rate of non-response because they are more difficult or sensitive. User convenience thus justifies imputation only to replace unknowns that have a negligible effect on the survey results.

Statistical organizations must establish guidelines for the judicious use of editing and imputation in order to take advantage of these techniques and, at the same time, avoid potential problems. The following are some useful recommendations:

(a) Emphasis should be placed on gathering good data, viewing the manual and computer adjustment as a backup measure that is unable to compensate for poor enumeration.

(b) The need for imputation should be kept to a minimum by careful preprocessing procedures and creation of realistic editing rules.

(c) "Not stated" categories should be included for items where the incidence of non-response exceeds a certain level, perhaps as low as five percent, depending on the variable in question.

(d) Users should be adequately informed of the changes that were made to the data during the course of editing. The rules for making these changes should be available in a usable form.

2. Stages of the edit procedure

The data cleaning procedure may be considered as consisting of a number of steps or "layers" of checks:

(a) Field editing and correction.

(b) Office editing.

(c) Machine editing of sample coverage and questionnaire format and structure.

(d) Range and consistency checks.

(e) Manual correction following machine editing.

(f) Automatic correction.

(g) Automatic imputation.

There is no "best" way to accomplish the task, and it is necessary to consider alternative approaches depending upon the circumstances. Factors determining the most appropriate approach in a given situation include the availability of personnel and facilities, the complexity of the questionnaire, sample size and volume of the data to be processed and the computer software being used. These considerations will determine the scope of each step, whether certain layers can be combined with others and the method of correcting errors. For each type of editing to be performed, the appropriate subject-matter specialist should write edit specifications and procedures for correcting errors, bearing in mind the available facilities and software.

(a) Field editing and correction

For a household survey of any complexity, scrutiny of questionnaires while the interviewers are still in the sample area is an essential requirement. Field editing permits access to the respondent for correction and additional information. At a later stage, once the questionnaires have been sent back to the office, it is rarely possible to recontact the respondent for additional information (except perhaps in longitudinal surveys involving repeated visits to the same set of respondents). Furthermore, only field scrutiny permits the discovery of consistent errors committed by particular interviewers in time for their retraining.

(b) Office editing

In most circumstances, it is necessary and cost-effective to subject the questionnaires to a further round of manual scrutiny in the office. It has two objectives: firstly, to correct major errors such as those relating to questionnaire identification; and secondly, to prepare questionnaires for coding and data entry so as to minimize the possibility of error in these latter operations.

Manual and machine editing are complementary operations, and the division of time and resources between the two needs to be optimized. In general, the smaller and more complex the data set, the more significant is likely to be the role of manual editing.

It is true that, unlike machine editing, manual editing does not permit complete uniformity of procedures and criteria for error detection and correction: fifty clerks may have fifty different ideas about how to resolve inconsistencies in the data. However, there can be a number of reasons in developing countries in favour of a more thorough manual editing. Firstly, it is often easier to recruit and train large numbers of office clerks than it is to enhance programming expertise and computer facilities. Secondly, while there are many packages available for data tabulations, and even for more extensive analysis, general purpose software for data cleaning purposes is not so plentiful. Finally, the complexity and relatively small volume of many household survey data sets are often arguments in favour of thorough manual editing prior to data entry and machine editing.

(c) Machine editing of sample coverage, and questionnaire format and structure

The first step in the machine edit procedure is to check that all questionnaires which are expected to be present are indeed present. For this purpose it is often useful to construct and computerize a sample control file on the basis of sample selection and implementation information, and to compare the questionnaire file with it on a case-by-case basis. Totals of number of cases by sample area may be prepared at this stage for subsequent operational control.

Next, each questionnaire needs to be checked for structural completeness, i.e. for the presence of all the necessary (and only the necessary) record types. This is especially important for surveys where more than one type or level of questionnaire is involved, or where complex multi-record questionnaires are present.

At this stage, questionnaires can also be checked for format. This will include checks on the presence of illegal or non-numeric characters and major column shifts during keypunching, apart from range checks on questionnaire identification and record types. Certain other transformations, such as conversion of blank (inapplicable) columns to numeric codes may also be convenient at this stage (World Fertility Survey, 1980).

(d) Range and consistency checks

The next step of editing looks at individual responses to determine whether or not they are valid codes. For example, if sex has a code of "1" or "2", an entry of "3" should be detected as invalid. This group of edit checks can also include a check for omissions. Some questions, such as age, may be obligatory for all or part of the population and a failure to respond should be indicated.

Consistency checks are performed to detect inconsistencies among responses. For example, a 14 year-old boy reported to be employed as a physician represents an inconsistency between age and occupation. Another group which could be classified as consistency checks involves searching for unreasonable entries or variants from likely ranges of value. An example of this would be a reported food consumption that is implausibly high.

Careful specification of the detailed editing procedures and rules is extremely important. The subject-matter specialists have a critical task to perform in designing edit specifications. They must take into consideration the tabulation requirements, the capabilities of the software being used, and all of the other criteria previously mentioned in determining a methodology for editing and correction of the data. Most importantly, they must communicate the specifications to the data processing staff in a format that is clear and precise and has the following features:

(i) An ability to specify the universe for each check being made; that is, the group of records or questionnaires to which it is applicable.

(ii) A clear indication of question or source code number for each item involved in the check.

(iii) A clear indication of the order and logical structure of the checks being made.

(iv) Clear conventions and notation for representing checks.

(v) Clear instructions for action to be taken upon success or failure of the check; if a message is to be written, the text of the message.

(vi) Specification of statistics required to indicate the type, frequency, and distribution of errors found and corrections made.

(vii) Provision for a verbal explanation of the check being made.

To facilitate the specification of edit checks a diagramatic representation of the questionnaire can be a very useful tool. The diagram may show the questionnaire structure, the valid codes for each question and the conditions under which it is applicable to particular respondents (World Fertility Survey, 1980). It is often much easier to see the interrelation of questions and edit rules when the editing is depicted in the form of a flow chart.

The important point is that, whatever format is adopted, the subject-matter specialist and the computer specialist must agree that it is mutually acceptable and fully comprehensible. If careful thought is given to writing the specifications, the time needed for the development and testing of the edit programs can be minimized.

(e) Manual correction following machine editing

Identifying errors in the data is the first step in editing. The data are considered edited only when the errors have been corrected. Corrections may be made manually and/or automatically by the computer. The approach adopted must take into consideration the nature of the error, precision required, analytical objectives, availability of personnel, capabilities of the software, effect on the schedule, and cost.

In the cleaning of household survey data, the general practice in many developing country circumstances is to use the computer only to locate errors, but to make corrections manually. There are several advantages to manual correction. First of all, it is extremely desirable to make corrections by referring to earlier stages of the operation: the data entry, coding and ultimately to the questionnaire itself. This can only be done manually. Secondly, rules for correction, particularly imputation, are likely to be more involved than edit rules for detection, especially for complex household survey questionnaires. Further, software with automatic correction and imputation facilities are more difficult to develop or acquire. However, serious limitations of manual correction should also be recognized:

(i) The process requires searching out of questionnaires and coding sheets, and can be extremely time-consuming.

(ii) It can be very difficult to keep a track of the number and type of corrections made. Careful organization of the way corrections are done is essential. Questionnaires should be easily accessible and located on shelves with clear labels indicating the survey round and sample cluster to which they belong. The editing staff looking up the correction must be thoroughly trained on how to interpret error listings from the computer, how to look up appropriate correction, and how to fill out the update forms.

(iii) The most serious problem can be the lack of uniformity in the way corrections are made. Human judgement is involved and unless stringent guidelines for error correction are provided and enforced, the data may become biased by personal opinions of individual clerks. The possibility of introducing new errors in the data also exists.

(f) Automatic correction

Automatic correction avoids the above problems, since:

(i) There is no need to search out questionnaires.

(ii) The computer can make the corrections much faster.

(iii) The computer will make a correction the same way each time.

(iv) Complex statistics on errors encountered and corrected can be maintained.

The appropriate strategy may be a judicious combination of manual and automatic correction. While this approach still necessitates locating questionnaires, it limits the number of errors that require access to the source documents for resolution. By contrast, imputation as such is a more purely computer operation. This is discussed further in the following paragraphs.

(g) Automatic imputation

Thus far, types of editing and general alternatives for correction have been presented. The discussion will now focus on methods of making corrections beyond returning to the field and simply rectifying data entry errors.

Errors can often be resolved by considering other data in the same questionnaire and imputing a response based on that information. For example, the marital status of a person who reports "relationship to head" as "spouse" can be corrected to "married" if it is in error, or "literacy" can be corrected based on "number of years of school attended".

When errors cannot be resolved in this manner, there is the choice of allowing the error to stand or relying on some other means of imputation. The simplest solution is to nullify the response by giving it a special response code which signifies "not reported" or "reported incorrectly". For example, if income is unreported, a code for "income unreported" can be created and assigned to this case. However, as was mentioned in a previous section, analysts may prefer not to see such categories in the tabulations except in cases of a moderate or high proportion of error. Other methods exist for assigning actual values.

One approach is the so-called "cold-deck" procedure, whereby missing or erroneous responses are replaced on the basis of a distribution of known cases. For example, errors in "sex" can be corrected by assigning "male" and "female" to alternative cases, since there is a known (generally 50-50) distribution. However, unless reliable data are available from previous censuses, surveys, or other sources, this technique necessitates pre-edit tabulation of valid responses from the current data, which may not be economically or operationally feasible (United States Bureau of the Census, 1979, Part A, p. 119).

Another approach is the so-called "hot-deck" procedure. In this case, referring to the missing income illustration given above, the income value reported for each person in a given occupation group may be stored into cells of a matrix. As a person with unreported income is encountered, he or she is assigned the income value for the last known case in the same occupation group. The outcome is similar to that in the "cold-deck" procedure, but current information is used in the allocation. This procedure is most effective if a certain degree of homogeneity exists between contiguous records or records that are grouped together possibly because of previous sorting. The effectiveness is improved if the replacement is selected on the basis of a match on characteristics which are highly correlated to the characteristics being imputed. For example, level of education might be used, instead of or in addition to occupation, to impute income.

One can easily see that if automatic correction is employed, the order of editing is extremely important. A priority of items must be established such that once an item has been edited it remains untouched. Otherwise, the correction procedure can become circular and the data can be greatly distorted. This is

particularly important if the survey involves a number of different questionnaires or when complex questionnaires with multiple records are involved.

3. Conclusion

The objective of any editing or imputation must be to enhance the quality of the data and to make it more convenient to use in tabulation and analysis. Excessive correction and imputation can distort the original data. A statistical organization must establish guidelines for the judicious use of editing and imputation. Emphasis should be placed on getting good data, viewing data adjustment as a backup measure that is unable to compensate for poor enumeration. The need for imputation should be kept to a minimum by careful preprocessing procedures, for example by regular, thorough and timely checks of interviewers' work while they are still in the field and by adopting realistic editing rules. "Not stated" categories should be retained, unless they are negligibly small to begin with (say fewer than five percent of the cases). The important thing to remember in editing, regardless of the approach taken, is that the objective is to present the truest picture of the universe represented by the survey and not to hide deficiencies in the data collection operation. The editing process must be given careful consideration, edit checks and correction procedures defined in great detail, and their application fully documented and controlled.

Even in a survey involving only moderately complex questionnaires, data editing and imputation can turn out to be a major task. Though editing and imputation rules may be determined largely by substantive considerations, the important questions to be considered by data processors are, first, the perfection to which it is economical to edit the data, given the rapidly diminishing returns from one "cycle" of correction to another; and second, the manner in which the task should be divided between manual and machine operations.

Generally, both manual and machine editing involve similar checks on the data, though the latter can be in much greater detail. Neither operation is dispensable, though they complement each other. Computer editing is faster, more thorough and objective, while manual editing is essential at least to remove gross errors and to prepare questionnaires for coding and data entry operations; also, its timing is more likely to permit the return of seriously deficient or incomplete questionnaires to the field. The appropriate division between manual and machine editing is an empirical question and depends upon a number of factors:

- The more complex the questionnaire, the more difficult and time-consuming it is to develop computer programs for the detailed edit checks; this favours manual editing.

-Larger sample sizes tend to make computer editing more cost-effective.

-The same is true when essentially the same questionnaire is repeated from one survey round to another.

-A most important factor is the availability of computer facilities, particularly suitable software and trained personnel. The less developed the facilities, the more appropriate it is to emphasize thorough manual editing.

-Even with machine editing, i.e. with automatic detection of errors, the corrections may be made manually or by the computer. Insofar as it is desirable to go back to the source questionnaires for correcting errors in the data, manual correction is preferable. The ideal procedure for making any correction or imputation is to verify first that the data on the file correspond to entries contained in the questionnaire, i.e. to verify that the error did not arise during the coding or data entry stages.

Finally, an interesting alternative to the more conventional approach to data editing described above may be mentioned. This incorporates both data entry and validation by utilizing interactive facilities. The possibility exists to build systems that integrate questionnaire coding, data entry, and data verification in an on-line environment. For each questionnaire, attributes can be entered either singly or by groups. At the end of each entry, plausibility conditions are tested and potential problems reported to the operator. While such a system may be more difficult initially to construct, it can lead to cleaner and more timely data since, once a questionnaire has been accepted, it is immediately ready for use in analysis (Sadowsky, 1980, pp. 19-20). The on-line editing approach presumes a much broader capability on the part of the data entry clerks, in that they must be able to rectify problems encountered. The approach does, however, seem an efficient one if the available personnel, hardware, and software can support it.

F. Recoding and File Generation

At various stages of the operation following data entry, it may be necessary to restructure the data set and generate new files, and to recode the existing data fields in individual records to define new variables more convenient for tabulation and analysis. With the development of computer facilities, microlevel data is increasingly seen as the final product of a survey. In continuing household survey programmes, linkage of data across surveys presents new potentialities and problems. These issues are discussed in the following subsections.

1. Flat versus hierarchical files

A file is described as "flat or rectangular" when exactly the same set of data fields exist for each respondent. For any respondent, the data fields are arranged identically within each record, and a fixed number of records with identical layout are involved. By contrast, a "hierarchical" or "structural" file may contain a different number or types of records for different responding units. In other words, the amount and type of data and hence the number and type of records may vary from one respondent to another. Hierarchical files may arise in a number of ways. For example:

(a) In a household income survey, different record types may be used to record the details of various major sources of income, one type for each source. The number and types of records present for any particular household will depend upon the sources of income enumerated for it, which will generally vary from one household to another.

(b) In a household interview two levels of data may be collected: data relating to household characteristics and data relating to individual members of the household, each with its own record type. The same household characteristic record(s) will be present for all households, but the number of individual member records will vary from one household to another depending upon the number of members in the household.

(c) In a multi-round longitudinal household survey, data from different rounds may be linked together. This may result in a structural file if for some households data only from some but not all rounds are present. Similarly, the linkage may involve different types of units, (hence record types) from different rounds, say household level data from one round and individual level data from another round.

(d) Similarly, in a multi-subject survey, different sets of variables may be enumerated over different subsamples, in addition to a set of core variables common to the whole sample.

In general, the processing of flat files is simpler than that of hierarchical files. Indeed, much available general purpose software requires data in the flat form. For this reason it is often desirable to convert hierarchical files to the flat form to perform specific processing operations. This conversion may be achieved essentially in two ways:

(a) By padding-in a lot of blank records so that the number and type of records involved for each unit is the same. For example, in the household income survey mentioned above, all relevant record types may be created for each household (blanks if necessary) irrespective of its particular sources of income.

Similarly, in the survey involving household level and individual level data mentioned above, the same number of individual level records (blanks if necessary) may be created for all households. By sufficient padding with blank records, any hierarchical file may be converted in principle to a flat file. In practice, however, this may not always be a feasible solution since the resulting file size may become excessively large.

(b) An alternative procedure in certain circumstances can be to split the original hierarchical file into separate files, one for each level of units. Each of the resulting files may already be flat, or can be converted into that form by padding with blank records.

A fairly typical illustration is provided by the World Fertility Survey. The basic arrangement consists of two types of interviews: a household interview in which basic characteristics of the household as well as demographic characteristics of all individuals in the household are enumerated; this is followed by detailed interviewing of women in the child-bearing ages. Either of the two interview files can be easily converted into a flat form by padding-in the required (generally small) number of blank records. However, the two interviews combined result in a hierarchical file. The issue as to whether it is better to process the two flat files separately or to process them together in the form of a single hierarchical file has been discussed in the following terms (World Fertility Survey, 1980, pp. 7-8):

"Data from the household schedule is sometimes punched separately from the individual interview data and sometimes together, household by household. Either way, the two types of data can be sorted together into one file or separated into two files as desired. A decision needs to be taken at the start on whether to process the data separately or together. ... The method chosen may depend on the software and hardware available. Some relevant considerations are:

-There is more likely to be software available for processing the separate files than the more complex combined file.

-If data are kept together and structure checked together each time data are read, errors of structure are less likely to be introduced when updates are made.

-Structure checking with two (separate) files implies matching them and identifying non-matches as one (additional) step in the process.

-Putting the data together may require sorting at an initial stage which may destroy the order in which data were originally punched. This in turn makes errors in the punching of identification fields more difficult to locate.

-With separate files, less data has to be handled at once which may be an important consideration on small computers.

-Using one combined file is conceptually tidier and involves less record keeping and a fewer number of computer runs."

The separate processing of individual surveys or survey rounds versus combined processing of merged (often hierarchical) files is likely to be a particularly important question for integrated programmes of surveys. This would apply not only to editing, but also to other phases of processing such as tabulation and analysis. In certain circumstances, with limited computer and software facilities, the appropriate approach may be to process individual survey files separately, and establish macro level links and comparison across surveys at the analysis and interpretation stage following data processing. In other circumstances it may be possible, and more economical, to combine data to be linked or compared into a single file for tabulation and analysis. Data linkage at the microlevel can greatly enhance the analytic possibilities. However, it should be recognized that such linkage can be a complex and time-consuming task; it is discussed further later in this section.

2. Addition of recodes

Once the survey data have been cleaned, it is often necessary to further transform or manipulate them to facilitate tabulation and statistical analysis. The process of defining new variables on the basis of existing data fields is called recoding.

For example, data from a household budget survey may consist of a large number of sources of income and items of expenditure recorded separately. For tabulation and analysis, it may be necessary to have these only by major groups. Several individual items may be combined to define new summary variables, and these recodes may be permanently added to the data file in order to avoid having to repeat the recoding procedure each time a derived variable needs to be referenced.

3. Microlevel data as the "final product"

Before computer processing was developed to its present level, government agencies conceived of their statistical output as the provision of specific tabulations, and subsequent data processing was confined to manipulation of the tabulated data. But this method tends to hide what may be important inconsistencies and differences within and among data sources, and to limit the use that can be made of the data.

With increasing computer facilities, the emphasis has been shifting from tabulations to the processing, editing, and storage of the primary or microlevel data. It is increasingly clear that data are most efficiently stored in the form of microunit records relating to each separate reporting unit. While the tabulations included in the publication programme of the survey are still considered the most important visible product of the survey, the survey "data base", from which the tabulations are derived, is more and more seen as a rich source of information available for a variety of unanticipated purposes in addition to the planned publications. This change in methodology permits the analyst more effective access to large bodies of information, at a relatively low and generally decreasing cost, and has made more feasible the relating of microdata sets directly to other micro as well as macro level data sets.

4. Data linkage

A primary objective of an integrated programme of surveys is to generate and analyze interrelated data sets from different surveys or survey rounds. Integration with data from other sources such as administrative records and censuses may be also involved. These analyses may explore relationships at the macro (aggregate) level, as well as at the micro (individual unit) level on the basis of linkage of data from complementary sources. The need to link data across different sources can substantially increase the complexity of the data processing task.

Special care must be taken in the design of survey procedures and questionnaires to ensure that such comparisons or linkage will be possible. Any comparison or matching on specific variables requires a uniform definition of terms, phrasing of questions, and coding of responses. When considering comparability with prior rounds of the survey, a question which always arises is whether it is preferable to introduce new and improved concepts or questions in a current round or to lean in the direction of comparability with the past. One solution often offerred is to proceed with thé improvements but also attempt some linkage with the past, perhaps by repeating the relevant questions as asked previously and then asking for the additional or amplified information. Where it is not

feasible to use both the old and new approaches in all cases in a survey, for budgetary or other reasons, there is the option of repeating the old concepts or questions for only a subsample, but one of sufficient size to provide for reasonably reliable estimates of the differences between the two procedures. Where the linkage data are sufficiently reliable, they can sometimes be used to revise the data from previous rounds in order to create a continuous series (United Nations, 1980a, pp. 23-24).

The objective of linkage of different data sets is to enrich the information available or to fill-in gaps in any particular data set. Linkage may be established on the basis of common substantive variables or on the basis of enumeration of common units or on the basis of a combination of these. The first, for example, would be the case in a survey programme in which individual surveys all incorporate a common set of "core" variables but are based on different samples of enumeration units. In the second case, different surveys employ common units of enumeration whether at the sample area, household or individual level.

A linkage of records from two or more files containing units from the same population is termed a "match". An "exact match" is one in which the linkage of data for the same unit (e.g., person) from the different files is sought; linkages for units that are not the same occur only as a result of error. Exact matching requires the use of common identification numbers. A "statistical match" is one in which the linkage of data for the same unit from the different files is not sought or is sought but finding such linkages is not essential to the procedure. In a statistical match, the linkage of data for similar units rather than for the same unit is acceptable and expected. Statistical matching ordinarily has been used where the files being matched were samples with few or no units in common; thus, linkage for the same unit was not possible for most units. Unlike exact matches, statistical matches are made on the basis of similar characteristics, rather than unique identifying information. Statistical matching is a relatively new technique which has developed in connection with increased access to computers and the increased availability of computer microdata files. As mentioned above, in a statistical match, each observation in one microdata set, the "base" set is assigned one or more observations from another microdata set, the "non-base" set; the assignment is based upon similar characteristics. Usually the observations are persons or groups of persons, and the sets are samples which contain very few, or no, persons in common. Thus, except in rare cases, the observations which are matched from the two sets do not contain data for the same person. A statistical match can be viewed as an approximation of an exact match (United States Office of Federal Statistical Policy and Standards, 1980, pp. 1-15).

In recent years there have been a number of efforts directed at statistical matching in certain developed countries such as the United States and Canada. However, there are several inherent

limitations of the technique: it may be suitable only for fairly "dense" data sets, i.e., when the sets being matched have a sufficient degree of overlap in regard to the common variables used for matching. Where there are only a few cases within broad matching intervals, the possibility of mismatching is obvious. For this reason, this matching technique is not generally applicable to records contained in small samples, or to those records in large samples which have unusual or extreme characteristics (Ruggles, et al, 1977, pp. 416-417). Above all, it should be emphasized that at present little is known about the nature and extent of errors present in data resulting from statistical matching. It is necessary to be cautious in the use of the technique. For these reasons statistical matching is not a satisfactory substitute for exact matching, and has rarely if ever been used when exact matching is possible (United States of Federal Policy and Standards, 1980, p. 32). In any case, the technique is inherently unsuitable in many situations and exact matching is called for. For example, if one wanted to compare the earnings of persons who had a given training programme with those who had not, an exact match between a list of trainees and earnings records would be needed. A statistical match between these two files would not be useful unless the earnings observations could be separated into persons who had been trained and persons who had not.

Exact matching is a more certain method of microlevel linkage of data. Errors in exact matching can be studied and their effect estimated in many cases. An example of fairly successful use of exact matching in a developing country is the linkage of data from the Intercensal Population Surveys (SUPAS) in Indonesia. The SUPAS II survey collected complete household information. In SUPAS III, 9,000 of the married females of child-bearing age from the SUPAS II sample were interviewed to collect fertility information. A computer match of the resulting files on the household identification successfully linked 90 percent of the cases; however, the remainder of cases had to be resolved manually because of identification problems. Characteristics of the villages involved as reported by the village headmen were then added to the resulting matched file. This three-way match posed no insurmountable problems and produced a rich data source for further analysis.

However, there are certain difficulties inherent in achieving exact matching. For example, in the United States, files of tax returns, social security records, and the Current Population Survey, have been linked with each other by matching the social security numbers which were reported in all three files. However, there were a substantial number of non-matches or mismatches due to non-reporting or errors in reporting of the social security number. Attempts to match files by using names and addresses of the respondents met with much greater difficulties due to the variation in names recorded in different files, the existence of duplicate names, changes in addresses, and even changes in names, e.g., following marriage. Thus, even in those instances where it is technically feasible, exact matching is costly to carry out. Above

all, exact matching is of course possible only on the basis of identical samples of units. In continuing survey programmes, it is generally necessary for various practical reasons to renew the sample, at least in part, from one survey to another, ruling out the possibility of exact matching on the full sample.

Regardless of whether exact or statistical linking is used, there are several problems common to all work in this area. These include data comparability, missing data, specific techniques for data linking and data manipulation, and the definition and evaluation of "goodness of a match" (Okner, 1974, p. 348). When matching is being considered, it is useful to assess whether it is in fact the best method of achieving the purpose. In some cases, the direct collection of data or some imputation technique, for example, might be better. As a minimum, the following factors should be considered in choosing the best method, giving each factor the appropriate weight for a specific application (United States Office of Federal Policy and Standards, 1980, p. 33):

(a) Amount of error in the results.

(b) Resource cost.

(c) Time required.

(d) Confidentiality and privacy considerations.

(e) Response burden.

The techniques described above offer the possibility of combining survey data with complementary data bases in order to increase their potential usefulness. It is clear that careful consideration must be given to determining the appropriate technique based on the nature of the data bases to be linked and the factors listed above. Countries may find that although the idea of complex linkage on a periodic basis is intuitively appealing, the cost and resource requirements are prohibitive and permit linkage only on an *ad hoc* basis at best. In relation to data processing the crucial question is the increased workload and complexity any particular approach might involve.

G. Tabulation and Analysis

1. Tabulation

For many descriptive surveys, the main output is in the form of cross-tabulations of two or more variables. Ideally, the general tabulation plan would have to be devised at the questionnaire design stage. However, at the stage of implementation, the data processor

requires from the subject-matter specialist detailed and unambiguous specification of exactly how each table is constructed and what its layout is. This should include:

(a) Specification of the date file(s) to be used.

(b) Specification of the variables to be cross-classified, indicating the specific variables defining rows, columns and panels of the table for each variable the categories to be included.

(c) The population to be included in the table.

(d) The statistics to be shown in the table, for example, frequency counts, row or column percentages, cell-by-cell proportions, means or ratios, if applicable, specification of the variable(s) used for the computation of cell-by-cell statistics.

(e) Whether the sample data are to be weighted or inflated.

(f) Table titles, subheadings, and footnotes.

(g) A sketch of the table layout, indicating details such as the size and number of cells, rows, columns and panels, and the cell entries to be printed.

Estimation procedures need to be worked out prior to the tabulation stage. For data collected on a sample basis with units selected with unequal probabilities, it would be necessary to weight the data appropriately before tabulation and analysis. (The appropriate weights are inversely proportional to the probabilities with which units were selected into the sample). Similar weighting may be required to compensate for differential non-response. Sample weights may be included as part of the data on each individual questionnaire, or may need to be added at a later stage by a matching procedure of some sort. It should be noted that the use of "non-self weighting" samples requiring weighting of the resulting data can be inconvenient in several ways: weights have to be computed, retained for a period and then used in programming and tabulation; their presence must be communicated to the future data tape users; and finally both weighted and unweighted frequencies would need to be shown in the published tables if they differ appreciably (Verma, et al, 1980, pp. 431-473). Weighting also tends to complicate the linkage of data across surveys.

Prior to running the tabulations for all regions and other major trends against the entire data file, it is essential to run the table programs on a test basis and have them verified for accuracy, format and presentation by the users. This can generally be accomplished very satisfactorily through using a sample of the data file and running the national tables only, and once these are verified, then producing the full set.

If large numbers of tabulations are to be run, it may be advisable to divide them into batches of tables in order to stay within the constraints of the tabulation software and avoid computer runs of long duration which tie up the computer and are more prone to be interrupted by equipment failure. Convenient grouping of the tabulations might be by subject-matter or by section of the questionnaire.

Occasionally, it may suffice for certain limited purposes to use the data at the aggregate level (e.g., at the level of the sample area, or by demographic, social or economic groupings of the population) rather than at the level of individual household or person. Such aggregation will obviously conserve space in the computer and allow faster access to the data.

If interactive computing is a possibility and the classical sequential file structure by questionnaire record(s) slows down the processing, the use of a non-sequential structure might be considered. The use of a transposed structure (i.e. sorting of data by variable rather than by interview) often allows survey data to be configured in such a way that ad hoc tabulations can be extracted from the data in a very short period of time using the appropriate interactive, user-oriented tools. The theory and use of such file structures is discussed in relevant Statistics Canada working documents and in the book, Time-Sharing Computation in the Social Sicences, by Edmund D. Meyers, Jr. (Sadowsky, 1977, 1978).

2. Computation of sampling variances

Estimation of sampling variances is required for interpretation of the data, as well as for more efficient design of future surveys. This latter consideration is particularly important for continuing programmes of surveys. With the availability of suitable software (see Annex I), routine and large scale computation of sampling errors may present no special difficulties. However, it is important to ensure that the necessary information on the sample structure is available on the data file to make these computations possible.

The items for which sampling errors are required, the frequency with which they should be computed and the methodology of estimation are concerns of the sampling and subject-matter specialists. Of course, these requirements should be kept within manageable limits. For example, where the same subject-matter is repeated overtime, it may be sufficient to compute the variances only periodically rather than on each survey occasion.

3. Other analytical statistics

The need may arise for other outputs to be generated in carrying out the analysis of the survey, such as regression coefficients, correlations, and derived indices. Care must be exercised to insure that the software, which is either specifically developed for these purposes or adapted from existing packages, is compatible with the sample design of the survey. This is especially important since the sample design for a continuing household survey is likely to be complex, so that available computer software applicable only to simple random sampling methods would not be appropriate for the production of analytical statistics. Furthermore, the file may need to be restructured in order to use relevant software.

H. Data Management

Any file or group of related files can be thought of as a data base. For example, a household survey might generate a main file of demographic information collected in the core questionnaire and several related files of the data collected in the modules attached to the core questionnaire. How these data are accessed depends to a large degree on the type of analysis being done and on the capabilities of the available computer system. Sequential processing may be quite adequate to produce the desired output. However, some types of analysis are best supported by a system which allows greater flexibility in looking at the data, such as easily being able to handle related subsets of the data base. Many of these applications are best served by a data base management system (DBMS); that is, a computerized system consisting of numerous components which have as their collection purposes the implementation, management, and protection of large bodies of data.

If the analysis to be done warrants the consideration of more than conventional processing and the computer system is adequate to support the storage and overhead requirements of a DBMS, there are a number of arguments in favour of establishing a DBMS:

(a) It promotes a degree of data independence whereby data definitions are centralized and independent of applications programs. This alleviates the need for extensive program modification and recompilation.

(b) Redundancy is reduced by not having to keep multiple versions of the same data set.

(c) Inconsistency is avoided by not having files which are in different states of update.

(d) Data can be shared. The data base can be manipulated to meet the needs of multiple users.

(e) Security can be enforced by limiting access to the data base.

(f) Concurrent processing is supported by allowing multiple users to access the data base simultaneously.

(g) The need for extensive sorting is eliminated by drawing on data structuring techniques.

(h) Non-programmers can access the data more easily.

Some, if not all, of these points could be effected without a DBMS; however, a DBMS minimizes the effort involved.

Each country attempting to implement an ongoing household survey must give careful consideration to managing the data it collects, taking into account the degree of file linkage and other complexities which might influence the need for data management software. Examination of organizational needs certainly involves looking at deficiencies in current data management. Lest the implementing organization view DBMS as a panacea for all data management problems, there are a number of issues which need to be taken into account. Experience shows that these considerations are often underrated or neglected at great future cost:

(a) The impact of a DBMS on an organization is disruptive.

(b) The technology of DBMS is new and difficult and requires substantial investment in training.

(c) The data base approach cuts across traditional installation management and requires staff reorganization and hiring.

(d) The transition to a data base system is highly visible, in particular because users outside the data processing department are inevitably involved in the reshaping of data needs and goals (Ross, 1978, p. 18).

The successful generation and implementation of a data base calls for a carefully organized plan. The first and most important step in data base generation is to determine the data organization requirements. This involves contacting the users of the data and gathering the information necessary to develop a data dictionary, a system of keys or unique identifiers, and a series of relationships which must be provided for. The second step is to identify the data processing requirements. These include source and frequency of

updates, type and frequency of updates, type and frequency of reports needed, and security or confidentiality requirements. Current as well as future needs should be considered. The result of this step is a list of all transactions and their characteristics, identifying the data base entities and relationships they involve and a sketchy outline of the data access. The third step is to generate structural definition of the data base. The nature and type of the data base transactions will often influence the particular hierarchy chosen.

It is generally accepted that it is unwise to attempt development of in-house software that matches the capabilities of DBMS. This consensus is based on the cost of both initial investment and continuing support. The vendor provides an important service in being responsible for maintaining DBMS software. Although in-house I/O modules and data managers are by no means a thing of the past, there is some feeling that these types of projects are not for any but the best data processing groups. An additional consideration is the time needed to implement an in-house system; DBMS and DBMS packages can be installed rather quickly and at a relatively fixed cost.

Implementing a data base management system may well be beyond the scope of many of the countries conducting household surveys for the first time. These ideas are merely presented to illustrate tools available to those countries which choose to pursue linkage of data and have the need to manage complex files in order to meet their analytical goals.

I. Other Operations

It may be possible and useful to computerize other survey operations, particularly in the context of continuing programmes where certain operations are repeated from round to round.

One such area is sample design and selection. A number of countries, including developing countries such as Republic of Korea and Kenya, have computerized the sampling frame along with the auxiliary information required for sample allocation and stratification, on the basis of which the required sample for each survey round can be selected easily and economically. However, for computerization of sample selection to be worthwhile, it is necessary that the sampling units involved are stable over a reasonable period of time so that the same sampling frame is usable without extensive revision for a number of survey rounds.

III. ORGANIZATION AND OPERATION CONTROL

The previous chapter discussed some important technical

considerations in the design of data processing procedures. This chapter deals with organizational considerations and the operational and quality control measures required for successful implementation of the data processing task.

A country planning to undertake a continuing programme of surveys will need to consider certain central issues relating to capability building in the area of data processing, such as: where and by whom processing will be done; whether the processing facilities should be centralized or decentralized; what are the staff development and training needs; and to what extent it is necessary and feasible to upgrade existing hardware and software. These broad issues will be considered in Chapter VI. In the present chapter various operational considerations are discussed, assuming that the basic orientation and organizational arrangements as well as the scope of the task to be performed have been determined.

A. Resource Planning

1. Budgeting for data processing

One of the areas most critical to the success of any survey programme is a careful study of how much it will cost and the ability to stay within the projected budget. The data processing task cannot be made simply to conform to budget dictated by other considerations; rather, it must be arrived at by a consideration of all of the individual components of the task, considering alternative approaches where possible. The consideration of alternative arrangements is particularly important when a substantial upgrading of the data processing facilities is sought, as often is the case in countries undertaking a regular programme of surveys for the first time.

The data processing budget should include provisions for indirect and overhead costs, inflation, as well as reserve funds for unforeseen costs. It should be based on the detailed data processing plan with carefully established estimates of workloads, rates of production, and personnel and related training costs. The budget should include the costs of all equipment, facilities, and supplies. Too often only large items of expenditures, such as data entry, are included in the budget while general office supplies are omitted (United Statues Bureau of the Census, 1979, p. 254). The development of an adequate cost-reporting system is crucial because budgetary estimates for subsequent years should be based on previous experience. Even though no two operations are exactly alike and circumstances change even when the same surveys are repeated, there are usually enough similarities with prior operations on which to base reasonable current estimates (United Nations, 1980b, p. 66). Unfortunately, very little information is available from existing household surveys on cost breakdown by specific activities such as

coding, data entry, editing and tabulation or on data processing costs in relation to other survey costs. Such information can contribute to a more efficient survey design and planning. Countries undertaking regular survey programmes should try to compile such data for their own benefit as well as for the benefit of other countries.

2. Creating a realistic schedule

Developing a realistic schedule is as difficult and important as arriving at a realistic budget and cannot be dictated by wishful thinking. Preparation of a calendar of activities can actually go hand-in-hand with the budgeting process. In both cases, a first step is to spell out all of the activities entailed in processing the data (United Nations, 1980b, p. 72).

These activities are first presented in the format of a system design, showing the interconnection of tasks. Each task is assigned a time estimate. Some activities, such as manual coding and editing, can be accelerated by increasing the number of persons working on them, whereas other tasks are limited by the availability of equipment or the ability to increase staff size. In the preparation of a calendar of activities, great care should be taken to make sure that the activities are arranged in the proper sequence and that realistic workloads and production rates have been determined for each activity. Tentative starting and ending dates have to be attached to each significant activity. It is useful to prepare the time-table in the form of conventional bar charts.

Major survey activities are interrelated; many activities cannot start until another activity is finished, or at least is underway. For example, the date to begin production processing cannot be before the date by which at least some completed questionnaires can be expected from the field. Once the date for receiving the questionnaires is set, planners can work backwards to set the dates by which processing plans and procedures must be completed, personnel hired and trained, programs developed and tested, and space and equipment made available. Working forward, the completion of the processing operations must be accomplished well in advance of the dates for publication of the results.

Similarly, the date by which the publication of results should take place determines the date by which manual editing and coding, keying, computer editing, and tabulation must be completed. In addition, these operations must be completed well before the planned publication date so that sufficient time will be available for review and analysis of the data prior to printing final reports (United States Bureau of the Census, 1979, pp. 258-259). The ongoing calendar of activities must conform to the schedule

established for rounds of the survey, thus avoiding a backlog of survey data. Time between a particular task in one round and the same task in the next round ideally should not exceed the time between rounds.

Network analysis, such as Critical Path Analysis, can be used effectively in determining a realistic schedule. Such techniques graphically depict relationships between activities and show the minimum time needed.

The relationship of the calendar to the originally targeted completion date should be examined realistically to see whether the initial time-table can be achieved. If that target appears to be unrealistic, it is better to face that situation at the outset than to suffer serious disappointments later. The estimates given for data processing are usually perceived to be excessive by those not familiar with complexities of the task. However, overly optimistic estimates at the planning stage give rise to unrealistic expectations; when the time actually taken exceeds the estimates, the data processors inevitably take the blame (Rattenbury, 1980, p. 4). It is preferable, for the sake of success of the survey programme and for staff morale, to offer realistic estimates based on past experience, instead of assuming optimum conditions that never existed.

The most efficient way in many circumstances is to plan the various activities to overlap each other. For example, office editing can begin as soon as sufficient numbers of questionnaires begin to be received from the field; coding and then data entry can start as soon as a batch of questionnaires has been manually edited. Such an arrangement is particularly desirable for surveys with long field work duration, and almost essential for a continuing survey programme.

B. Organization and Management

Following the estimation of the overall budgetary requirements and time schedule, the planning process needs to focus on specific requirements for implementation. These concern staffing and lines of communication, equipment needs, space requirements, and plans for management of the activities of the computer facilities.

1. Organization and staffing

The data processing staff should have representation equal to that of the field staff, sampling staff and analysis staff, in the overall management established for the survey programme. The

organization of the data processing staff itself is to a large degree dependent on the existing personnel structure. However, the following positions or groups need to appear somewhere within that structure: director of data processing, computer centre manager, operations staff, clerical coding and editing staff, operational control unit, data entry staff, and systems analysis and programming staff.

For processing the household surveys, a system analyst should ideally be assigned as the data processing manager, and a group of programmers permanently designated to provide programming support. It is desirable to build up the staff on a permanent basis in order to ensure continuity and avoid constantly having to train new staff.

The various groups involved in the survey processing should be co-ordinated through the processing manager. Within each group there should also be a supervisor who will assume responsibility for that area of processing.

The persons available and their respective levels of skill should be matched against the requirements dictated by the scope of the task and the calendar of activities. It may be necessary to hire new staff and to provide training to existing staff; these should, of course, be provided for in the survey budget.

Potential candidates to fill new positions should be thoroughly interviewed and their credentials examined. Evaluation factors include: prior education and experience, interest in the type of work to which they would be assigned, their career goals, willingness to make a commitment to the job, ability to work well with others, and recommendations from previous sources of employment. It is important that new staff be able to work in harmony with existing staff.

2. Training needs

Training needs vary greatly from country to country. In situations where data processing facilities are newly established or substantially upgraded, an intensive and long-term programme of training will be needed.

If new computer equipment is acquired, the operations staff will need training in order to use it properly. This is also true of data entry equipment. A few days of intensive training for office coders and editors is essential before each new survey, with full discussion among all participants of any problems that come up during the coding and editing of sample questionnaires. Data entry

operators also need a short training course to show them how to deal with identification fields. To help motivation in all of these areas, the training should include a brief exposition of the purpose of the survey programme and should make the staff feel that they are involved in an important project (Sadowsky, 1980, p. 8).

Systems analysts and programmers should have or gain experience in programming in at least one higher level language to be used in the system. If generalized packages are acquired for editing, tabulation or analysis, comprehensive training in their use should be provided to all the staff involved. Analysts and programmers should be thoroughly acquainted with the guidelines for programming and testing so that all work proceeds in a uniform way.

On-the-job training at all levels is an essential requirement for the successful implementation of any survey programme. A programme of "cross-training" should also be developed to familiarize subject-matter specialists with data processing concepts, and data processors with survey concepts. Many problems are created through ignorance of how certain decisions will affect the work of others. A small investment in cross-training should have a significant effect on morale and should pay off in a more efficient survey operation.

3. Lines of communication

From the beginning, clear lines of communication should be developed and maintained. Periodic meetings should be held at which representatives from the various data processing groups can exchange ideas and report progress. When necessary, these persons should be included in higher level meetings to discuss problems or explain advances in their areas of responsibility. A formal reporting system, co-ordinated by the data processing manager, should be developed whereby a weekly or monthly status report reflects the status of each data processing activity.

4. Equipment

It is necessary to estimate the requirements for various types of equipment such as computer hardware, data entry equipment, adding machines, calculators, typewriters, duplicating and printing equipment, and miscellaneous office supplies.

The quantity of data entry equipment available for converting the data to machine-readable format should be evaluated against the anticipated volume of work and the time frame in which it must be accomplished. It may be necessary to acquire additional data entry devices in order to meet the schedule.

In relation to computer hardware for processing, two important factors to consider are the adequacy of the hardware to do the job and its availability for the household survey programme. In order to properly assess these issues, the requirements of the software to be used must be known and an estimate of computer time required must be attempted. The amount of computer time required to perform all editing, correction, file structuring, and tabulation is quite difficult to estimate. Much depends on the quality of the recorded data, the appropriateness of the overall data processing plan, the speed and availability of the computing system, the software used, and the error (human and other) incurred in processing. The adequacy of the hardware should be judged in the light of its ability to support the software to be used. There may be requirements of minimum memory size, necessary peripherals, or particular compilers. The processing capabilities of the machine should be matched against the estimated computer time and the desired turn around time to determine if the system can do the work on a timely basis.

Equally important is guaranteed access to the machine for the specific job. A detailed schedule of access to the machine should be worked out well in advance of actual processing, and should include time for program development and testing, installation of packages, acquiring of practicl experience in the use of new software, and production processing. Some extra time should be included for unforeseen problems. How the access is to be provided should also be discussed. For example, if programmers interact with the machine via terminals, there must be an adequate number of terminals to ensure efficient use of the programmers' time. One rule of thumb is that, in the ideal situation, at most two moderately active programmers or users should share one terminal, and an especially active programmer or user should be assigned a terminal for his or her exclusive use (Sadowsky, 1977, p. 21).

Evaluation of equipment requirements may dictate the need to acquire new equipment or augment existing equipment. The procurement process is generally very time-consuming and should begin well in advance of when the equipment must be operational.

5. Space considerations

The planning process should take into consideration the space requirements of the various processing activities. This includes assignment of space to individuals and equipment, environmental control of that space, quality of the electrical supply, and planning for storage of survey materials and computer tapes and disks.

A proper physical working environment is a necessary condition for an efficient data processing operation. Computers and data entry equipment require space that is controlled for

temperature and humidity. Punch card stock, in particular, may warp or otherwise prove to be unusable if subjected to frequent and excessive changes in humidity.

The physical environment for the clerical and data entry staff should also be considered. Planned production rates and quality levels will be difficult to realize if the physical arrangements are unsatisfactory. Sufficient space must be available to enable a smooth and steady flow of work to be maintained, including provision for temporary storage of questionnaires adjacent to the clerical and keying operations. Adequate lighting and ventilation are essential for good quality work (United States Bureau of the Census, 1979, p. 140).

The arrangement of the space is extremely important. Space should be allocated to minimize the movement of materials over long distances, especially survey questionnaires and other bulky documents. For example, since it is more inconvenient to move questionnaires than computer printouts, it is more important to locate all processing operations using questionnaire as close to each other as possible.

If new equipment is being procured or the existing computer site has experienced electrical problems, serious thought should be given to electric power considerations. The quality and nature of local electric power available for supporting a computer installation is an important factor in determining the ease and success with which the equipment can be installed on site. While data processing equipment varies in its ability to tolerate fluctuations in power supply, all equipment is affected adversely by these to some degree. Irregular power supply can cause unpredictable failures which can be confused with hardware or software errors. Clearly, it is important to understand the characteristics of power required for reliable operation of specific equipment, and to meet these requirements.

Two steps are generally necessary to provide appropriate power. First, the characteristics of the public power supply should be measured as accurately as possible, and the nature and extent of deviation from what is required should be determined. A power line disturbance analyzer used over the course of several weeks will provide the needed information. Second, power conditioning equipment must be provided to condition the existing power, generate independent power, or to combine the two approaches as in the case of uninterruptible power supply systems (Sadowsky, 1979, p. 16).

Adequate shelving and filing space must be provided to assure the orderly maintenance of survey materials, system documentation and output. Metal shelving is preferable for storing items for a fairly

long period of time. Shelving should be of a type that can be easily assembled, and the shelves should be adjustable. The shelves must be sturdy. Filing cabinets should accommodate computer printouts, since they are an essential part of the survey documentation.

The storage area designated for magnetic tapes and disks should be environmentally controlled and should allow orderly access to the data files. If the tapes are not stored in cans, wraparound jackets should be provided for the tapes in order to prevent the accumulation of dust on them which can interfere with successful recording and reading of data.

6. Management of computer facilities

A proper management of computer facilities is vitally important. This cannot be achieved by a casual attitude but demands constant attention and a dedication on the part of those in-charge. Without spending large sums of additional resources, the "computer centre" can maintain or institute procedures and policies that promote efficient utilization of resources and user satisfaction.

The computer centre should adhere to a standard schedule of operations which provides maximum use of the machine. Periodic preventive maintenance and regularly scheduled use of the machine for other work should be clearly stated.

The centre should be run according to a set of established regulations. It is important to permit access to the machine only to the operations staff, systems programmers, and those who have a legitimate reason for being in the machine room. The same regulation should apply to the data entry area, where unnecessary persons get in the way, impede production, and can damage the equipment. A priority scheme should be designed to maximize throughput on the machine. This can only be effective if no exceptions are allowed.

The above policies should be compiled in documentation distributed to every user to increase public awareness of the standard being followed. An informed user is less likely to err in following regulations.

The computer centre should make every attempt to be responsive to its users. Jobs should be handled as expeditiously as possible, making sure that output is promptly delivered. Persons should be designated to provide assistance when the user has questions. Changes in the operating system or new software should be announced in written form to the user and training should be provided if necessary.

An effective system for controlling tapes and disks should be established so as to avoid losing, misfiling, or writing over a final data tape or disk. One or more persons should be designated to manage this library in order to control it properly. Good communication within the computer centre and with its users can be enhanced by design of appropriate forms so that important information, such as which tape is for input and which for output, is not left up to the operator's interpretation.

Use of a job accounting system benefits both the computer centre and the user. The computer centre can use it to monitor utilization of the machine, issue bills to users, and plan for the future. The users have a record of their work and machine utilization.

Every effort should be made to maintain equipment in a proper condition and avoid unexpected periods of down time. There is nothing more frustrating to the user than being in the middle of a long run and having the machine crash because it is long overdue for preventive maintenance. It may be necessary to train computer centre staff in how to maintain the equipment and keep spare parts in reserve. In some instances, machines have been down for weeks because of a damaged part which had to be ordered.

C. Quality Control and Operational Control

The control systems applied to the processing have a major effect on the timeliness and quality of the data. For convenience in describing control measures, quality control is distinguished from operational control, although the two are closely related. Quality control refers to maintaining the quality of processes and data at an acceptable level. An operational control system refers primarily to the maintenance of uniform records of production, expenditures, and flow of materials through the various operations (United States Bureau of the Census, 1979, pp. 261-262).

1. Verification of office editing, coding and data entry operations

Quality control is most commonly connected with a verification system which checks the quality of an operation performed. In large-scale operations, it is important to be able to quickly remedy a situation which propagates the same error throughout the data or to identify a person who consisently fails to meet the necessary requirements of his job.

Verification of office editing, coding and data entry operation is an essential requirement of quality control. In any

verification process, the ideal approach is to have two persons perform the operation independently and then compare the results, because in "dependent" verification the verifier tends to agree with the work of the producer although it may be incorrect.

For editing this would imply that the editor-verifier should do essentially the same job as the original editor. Verification is not a matter of merely checking the cases where the original editor found errors, but of checking the whole questionnaire; i.e. verification should be done as if the original editing had not been done. This is not an easy proposition, considering that the corrections have already been written on the questionnaire.

When coding is done in the spaces provided on the questionnaire itself, independent verification can be accomplished in some instances by covering the space in which the code is placed and having the verifier repeat the thought process from the beginning without being biased by seeing the code supplied previously. Such is the case regarding data entry verification, where it is necessary to actually have two clerks enter the data and then match the files for discrepancies. Repeating the whole process can substantially increase the time and cost involved, and may not be possible in many circumstances.

Ideally, 100 percent verification is desirable at the initial stages of an operation, not only to correct errors but also to identify clerks with below average performance. Subsequently, verification on a sample basis should suffice in most circumstances. However, any clerk must first qualify for his work to be verified only on a sample basis (rather than on 100 percent basis) by demonstrating achievement of a certain level of performance. The performance should be monitored on a continuous basis; it may be necessary to increase the proportion of questionnaires verified, even to reintroduce 100 percent verification, for a clerk who fails to maintain an adequate standard of work.

2. Quality control of machine editing and tabulation

Computer editing and imputation must be used judiciously and must be thoroughly controlled in order not to introduce new errors in the data. During machine editing, each computerized checking run must always be rerun after corrections are made and this must be done repeatedly until no more errors are detected. This is often referred to as cycling the data. Shortcuts taken here to catch up on the schedule will mean that incorrect corrections go undetected, and in the worst case, the edited data may contain more errors than the original data. However, spending of a great deal of time on a negligibly small number of residual errors should be avoided (Rattenbury, 1980, p. 16).

An important part of controlling the quality of data during the machine processing will be by keeping a diary. A diary is a printout of information which is used to analyze the output of a given procedure or program to ensure that expected levels of quality are being maintained.

Quality control of the tabulations involves a review of the computer printouts for consistency within and between tables. In cases where the printouts are the camera copy for printing, each publication table should be reviewed to ensure that it has the correct title, correct headnote and footnotes, clear type for printing, etc. (ibid.)

3. Quality control of hardware and software

Hardware and software quality need to be assured well in advance of the beginning of the processing and, in the case of a recurring survey, must be constantly checked for the duration of the survey. Malfunctioning hardware can have a disastrous effect on the processing. The computer centre should establish and rigorously enforce a fixed routine of machine maintenance. Check lists for maintenance should be developed and test decks supplied by the manufacturer should be put through the computer at regular intervals to detect machine failure. Devices such as "skew tapes", which are available from some manufacturers, should be routinely placed on the tape drives to assure that the read/write heads have not gotten out of tolerance. Records should be kept on the types and causes of machine failure, the time required to repair the defect, and the date. These records should be analyzed periodically and action taken to eliminate the causes of machine failure.

It may be necessary to install one or more generalized software packages for editing, tabulation, or analysis. In each case, the package should be completely benchmarked and tested well in advance of the time when it is needed. A malfunction of a package written elsewhere can often be difficult to correct; it is desirable to detect such errors early.

As with other aspects of quality control, the goal is to make certain that the quality of incoming data does not deteriorate during the course of the computer processing. One device used to measure attainment of this goal is the trace sample. A trace sample consists of a small sample of fictitious cases that represent a wide range of situations. The data should be created to test every path in the program so that editing and tabulating procedures can be fully tested well in advance of having actual data available. The sample is followed through the computer processing and examined on a before and after basis to see whether the operations performed on the data by the software, hardware, and computer personnel were in fact done correctly (United States Bureau of the Census, 1979, pp. 191-192).

All parameters to generalized software packages and all custom-written software should undergo exhaustive testing. The trace sample is an excellent tool for testing. Beyond that, a sample of "live" data should be used just to see if there are any idiosyncracies of the actual data that might necessitate program modification.

The subject-matter specialist and computer specialist should identify milestones throughout the software development process at which feedback will be solicited from those who will use the data. These will serve as periodic checkpoints to assure that processing will meet the needs of the users and will produce statistically correct results.

4. Operational control

Operational control plays an equally important role. The primary purposes of an operational control system are:

(a) To determine the status of the work on a current and cumulative basis at any given time. Essentially, this involves the ability to measure output of each operation as well as to indicate the existence of backlogs and poor performance rates.

(b) To measure current and cumulative expenditures in terms of staff, time, and money for each operation.

(c) To ensure that the proper materials are processed through all applicable operations; for example, the flow of questionnaires, diskettes, tapes and disks should be controlled.

(d) To ensure the prompt transmittal of materials from operation to operation (ibid., p. 262).

Without an operational control system, a backlog or conflict in activities or an overrun of resources could seriously threaten the success of the survey programme.

The control of work progress is important to ensure that schedules are met. A master chart of all data processing activities must be carefully maintained and regularly compared to the proposed schedule to be sure that work is progressing as planned. Discrepancies must be resolved as they occur by adding resources, adjusting the schedule, or by some other means. Progress reports should be written on a regular basis by persons managing various

aspects of the survey. The composite report should accurately represent the status of the survey. As noted earlier, creating and adhering to a realistic schedule is essential.

There are several types of forms that are necessary for all operational control systems, including:

(a) Inventory forms.

(b) Transmittals to control the flow of materials through the various operations.

(c) Production records to control the progress of the work.

(d) Cost records to control expenditure.

Each control form should be designed for a specific use or operation. The designers of the control forms should consult with the users before the forms are finalized to ensure that the forms will ultimately produce the information required for control and reporting. The forms should be designed so that they can be completed quickly and easily by the control clerks (ibid., pp. 262-264).

5. Batching data for processing

Processing the data in batches or work units offers several advantages. First, the need for human and machine resources is spread out more evenly. Second, individual computer runs are shorter and not only fit between other jobs but are less likely to be aborted by machine or power failures. Third, the data processing up to the end of machine editing can often be finished more speedily (Rattenbury, 1980, p. 6). Last but not least, problems caused by incorrect specifications, misinterpretation of specifications, or programming errors will be less costly to detect and correct because a reduced amount of data is involved.

It is generally advisable to base the batch assignment on a geographic level or some similar characteristics easily identifiable on the questionnaire. In this way questionnaires, data, and control forms can be readily associated with particular batches.

The process of managing an orderly transition of many batches of questionnaires through the operations of establishing shelf storage, recording on diskette, verifying the diskette, transferring the data to a tape or disk file, cycling through error detecting and correcting procedures requiring one or more rounds of correction transactions, and adding the batch to the master file, is not

trivial. Without proper controls, batches can miss one or more processing steps, be incorrectly updated, or even be lost. It is necessary to have a system that reflects the status of an individual batch, as well as the overall production status, at any point in time. This can be done manually or by a straightforward computer program to which are posted all changes of status that occur during a processing period, such as a day.

There are some advantages to implementing such a system on a computer. Cross-indices, indices by different keys, and reports regarding progress to date are easily generated, as are projections of completion dates for all phases of processing. If adequate time and resources are available, this type of system could be implemented and should prove to be cost-effective (Sadowsky, 1979, p. 22).

D. Documentation in Support of Continuing Survey Activity

One of the key factors in the success of an ongoing survey programme is effective documentation. Too often documentation is considered as an afterthought, if at all. Without proper documentation, time and money will be lost and users of the data will become frustrated and may even stop supporting the data collection effort. Many statistical agencies have learned by bitter experience that inadequate documentation can result in loss of valuable information because stored data can not be processed for technical reasons, or can be processed only by incurring costs which could have been avoided if satisfactory documentation were available (United Nations, 1980b, p. 53).

Documentation should be elaborated as a part of the planning and implementation of each statistical project and in accordance with clear rules for the division of work. Moreover, the documentation should be designed in conformity with standards that are well described and easy to learn. Finally, the documentation should be maintained up-to-date.

To satisfy these requirements, the documentation must be elaborated and maintained by several units of the statistical agency in co-operation, namely, the subject-matter divisions, the systems and programming unit, the machine operations unit, and the central units that may exist for information and promotion of the use of statistics and for printing and storage of publications and questionnaires. However, standards for the documentation should be prepared centrally.

There are several levels at which documentation should be provided. Systems analysts and programmers, computer operators,

managers, and users of the data all have diverse but critical needs for documentation.

1. System documentation

The systems analysts and programmers will not only be called upon to design and implement the initial system of programs, but will be expected to maintain and enhance those programs as the need arises. Given the high turnover rate often found among this group of people, one can see the necessity of a well-documented system. This documentation should minimally include the following components:

(a) A system flow chart, which shows how the individual programs fit together to make up a system.

And on a program-by-program basis:

(b) Complete specifications for the program fully describing the inputs, outputs, and procedures to be followed.

(c) An up-to-date flow chart of the program.

(d) A well-commented source code listing of the program, indicating the author and location of the program.

(e) All test runs, indicating the test being performed in each case.

A programmer writing a program often feels confident of completely understanding and remembering every detail of code, but time has a way of erasing one's memory and as little as six months later, he may not remember what a particular routine does, or why. Any modifications made to the initial system should be reflected in the documentation. These items of documentation should be kept in a central file so that they can be easily accessed.

2. Operational documentation

Production running of the system will most likely be the responsibility of the operations staff. Although they need not understand the inner workings of the programs, they must understand the general purpose of each, how they fit together, and in what sequence they are to be run. The documentation prepared for the computer operations staff should include the following:

(a) A system flow chart, emphasizing the inputs and outputs of each program and the disposition of the outputs.

(b) Instructions for running each program.

(c) A schedule for production processing.

The more this group understands about the total system, the more likely they will be able to cope with operational problems when they occur.

3. Control forms

An additional area for documentation which may span the development and operations functions, or exist on its own, is that of control of materials. The control forms that are used to monitor the transfer of questionnaires, cards or diskettes, tapes, printed tabulations, and any additional materials should be maintained in a central location and serve as complete documentation of the entire processing effort (see section C.4 above).

4. Study of error statistics

Additional documentation concerning error statistics is often useful. It is important to understand what effect editing had on the data. The diaries that are produced during computer editing should provide information on the number of changes by question and the number of questionnaires having errors. An overall effect of editing can be obtained by comparing the data on a questionnaire-by-questionnaire basis before and after editing. This information can be used to identify weaknesses in and improve the questionnaire and the interviewing process. If matching of microlevel data is involved, statistics on error rates in matching can be usefully maintained.

5. Description of procedures

In addition to the detailed documentation described above, data processing procedures should be described in the reports on survey methodology. The data processing chapter should address the following topics:

(a) Basic decisions on data processing, including hardware procurement and packaged software acquisition.

(b) Receipt, check in, and operational control procedures.

(c) Manual editing, coding, and diary review procedures.

(d) Keying procedures.

(e) Computer processing procedures for coding, editing, and tabulation; methodology and extent of imputation.

(f) Quality control procedures.

(g) Personnel functions and requirements, including training activities.

(h) Budget, costs, and person-hours expended, by operation.

Documentation of procedures serves as a record of accomplishments, and can form an excellent base for planning future activities (United States Bureau of the Census, 1979, p. 140).

6. Guide for users

A guide should be developed to provide the user with the information necessary to understand and use the data without making this experience a source of frustration. The user documentation should include comprehensive description of the data including codes or categories for each variable, processing specifications at various stages, indication of data collection methodology and data quality, where the data are stored, and physical characteristics of the storage. These requirements are discussed in the following section.

E. Data Documentation and Archiving

The collection of survey data involves considerable cost and effort. Increasingly, their usefulness goes far beyond the basic descriptive cross-tabulations which may need to be produced as soon as possible after data collection. The storage of data at the microlevel for possible use by a variety of users and researchers necessitates detailed data documentation (the following paragraphs draw extensively on World Fertility Survey, 1980).

1. Data files

During the processing of survey data, a large number of different files are created. Once the data have been cleaned and restructured and recoded as necessary, the different files generated during this process should be reviewed. Files not required any more should be discarded and others fully documented and retained for future use. In general one should consider keeping at least three versions of the data:

-The original, uncleaned, raw data (after manual editing and correction, of course).

-The cleaned raw data.

-The data restructured and recoded for tabulation and analysis.

If systematic computation of missing data is involved, two versions each of cleaned raw data and recoded data are desirable, namely, the versions before and after imputation.

2. Code book

The basic documentation for the actual data is the code book. The code book specifies each variable that is in a data record, giving its location in the record, its name, description and meaning of codes including non-stated and non-applicable codes. It is similar to the coding manual used by office coders during the coding process, except that it need not contain any coding instructions. The code book must be prepared before starting to process the data by the computer. Its preparation is a useful way for the data processing personnel to familiarize themselves with the data.

3. Machine-readable data description

Machine-readable versions of the code book are extremely useful for analysis of the data. All general purpose data analysis software require a description of the data. This data description consists, at a minimum, of the location of the different variables in each type of record. More sophisticated packages provide for labelling of variables and variable categories. For this, the information has to be supplied in machine-readable form.

4. Marginal distributions

Any data analyst needs to know the distribution of the variables that are used for the analysis. Such distributions may be produced when required, but it is convenient to have them archived with the data for easy reference.

In fact the role of marginal (frequency) distributions is wider than merely the convenience of the user of the "final" data. At various stages in the data processing operation, the marginal distributions are a basic tool of monitoring the data at hand. They should be produced, for example:

-Before range and consistency editing and correction of the raw data, once the file has been edited for structural completeness and correct format. This gives information on the quality of the data, including an indication of the need for correction and imputation.

-After the raw data have been cleaned, to confirm that all values are now valid and to provide a reference document for the data in their original form.

-For any restructured and recoded files, to confirm that these operations have been correctly carried out, and to provide systematic information for the design of the tabulation and analysis plans.

-If systematic imputation is involved, frequency distributions should be produced both before and after imputation.

-In cases where the sample data have to be weighted or inflated before statistical estimation, it may be important to produce both weighted and unweighted frequency distributions. The unweighted distributions give the sample size of the various categories, which determine the sampling variability. The appropriately weighted frequencies give the relative significance of the categories in the estimations derived from the survey.

5. Survey questionnaires, and coding, editing and recode specifications

A copy of the original questionnaires should always be available to the user. Where codes are not given on the questionnaire itself, then coding instructions used and the detailed code should also be provided. Similarly, when new variables are defined, the recoding specifications should be documented. All specifications developed for processing the questionnaire should be updated to reflect exactly what was actually done; it is preferable to have the actual programs or control commands available. This is especially important for coding and recoding specifications where changes may have been introduced subsequently to the initial specification.

6. Description of the survey

For the analyst and user of the data, the code book and data processing specifications are not sufficient to provide a full understanding of the data. A written document containing special notes about the survey and the way it was conducted is also essential. The following information may be summarized and appended to the archived data:

-A statement of the nature of the data being documented, a list of the different files of data available and references to related documents.

-The name of the executing agency which carried out the survey.

-A description of the sample. This should include whether it was stratified, the number of area stages and number of clusters and whether it is self-weighted. If weights are used, then indication should be made of whether they correct only for unequal final probabilities or also for differential non-response. The rules by which the weights are assigned to the different respondents in the data should be given.

-A short description of the questionnaire. For recode files, any section of the questionnaire not in the file should be mentioned.

-Details of the field work and office editing and coding giving the numbers of people involved and the dates each stage took place. In addition, a short comment on editing procedures and a list of the edit checks used could be given.

-Data processing methods and software used for checking, correcting, imputing, recoding and tabulating the data.

-Imputation procedures used, a summary of which were imputed variables and for how many cases imputation was done.

-Any other information and peculiarities of the survey data collection and processing not noted before.

-The structure of the data file: whether it is "hierarchical" of "flat"; if there is more than one record per case, details of the different card types and whether they are obligatory or optional; the way the file is sorted.

-Explanatory notes on individual variables where there are, for example, known errors or deficiencies or where further explanation beyond that given in the code book is required.

7. Need for an in-depth manual to address the data processing task

Once the survey design including the questionnaire content and the tabulation plan is determined, it would be most useful for the national survey organization to develop an in-depth manual on data processing procedures. The manual should provide most of the users' documentation described above. It should list in detail the codes used, specify the various edit checks to be made and procedures for error correction and imputation, define how derived

variables (if any) are to be constructed from the raw data, specify each table in terms of the questions or derived variables used for its construction, list how other statistics such as sampling variances are to be computed, and define the microdata files to be created. The objective of the in-depth manual will be, first, to develop an understanding between data processors and other survey specialists of the problems and procedures to be used, and second, to document the procedures in sufficient detail to assist in the development of required computer programs and procedures and in the implementation of the data processing task.

In fact, such a manual represents a complete collection in a single place of all documents relating the data processing of the survey. During the data processing phase, it is the working document containing all necessary documents and specifications for the preparation of computer programs and for controlling the data processing. At the end of the data processing phase, it forms a complete record of all processing including the final specifications and listing of all programs (or parameter cards in the case of package programs) used. The Data Processing Guidelines developed by the World Fertility Survey, referred to above, provide an excellent example of such documentation. In fact it is instructive to list here the contents of the "data processing manual" recommended in the above mentioned publication:

(a) Data and documentation

- -Copies of survey questionnaire.
- -Data dictionary proforma (forms for recoding variable names, locations and codes).
- -Code book for raw data.
- -Machine-readable code book for data restructured or recoded as required for tabulation and analysis.
- -Test data.

(b) Planning and control

- -Data processing flow charts.
- -Programming and data processing estimates of time required.
- -Bar chart for the above.
- -Data processing control documents (indicating timing of individual steps by processing cycle, for each batch of questionnaire).

(c) Data processing specifications

-Data entry specification (card layout).

-Format check specification.

-Structure check specifications.

-Network diagrams for various questionnaires.

-Range and consistency checks.

-Restructuring and data recoding specifications.

-Specification for tables (in terms of variables).

-Specification for the computation of sampling variances (and other analysis).

(d) Specification of the programs used, indicating the purpose, inputs and outputs, flow chart and source of each program.

(e) Sample runs of programs.

IV. TRENDS IN DATA PROCESSING

In order to assess the level of data processing technology in a country and to consider areas for future expansion, it is useful to understand current trends in such areas as data entry, hardware, and software. This is not to say that developing countries should be striving to achieve the state-of-the-art situation in areas such as these; to the contrary, countries need to appreciate the alternative technologies in order to choose the mix which provides the most appropriate support to their particular applications. In some cases, the most recent innovations may be extremely effective in a business or manufacturing environment, but would not serve the needs of a national statistical agency. The following discussion of three areas in which there has been dramatic change over the years is offered in an effort to provide a broader base for evaluation and decision.

To indicate realistic directions of future development of the data processing capabilities of countries participating in the NHSCP, the next chapter will place these recent trends in the context of the existing data processing facilities and practices in national statistical offices in developing countries.

A. Data Entry

Computers have always processed data far faster than it has been possible to get data into and out of them. The progress in data entry techniques and equipment has been modest in comparison with the phenomenal gains in the rate at which data can be processed within the computer (Lusa, 1979, p. 52).

Data entry as most people recognize it today began in the late nineteenth century when a young engineer named Herman Hollerith invented the 80-column punched card. Taking a cardboard replica of the old United States dollar bill to ensure that his card was treated with respect, he cut holes in it to represent data. Automatic data processing systems were forced to live with the restrictions of the 80-column card for over a half a century. However, the manipulation of entered data became comparatively easy with the invention of the electronic computer. Little attention was devoted to input systems until the 1960's when processing speeds of "third generation" computers demanded increasing volumes of data (Aldrich, 1978, p. 32).

The first technology update of the old card-cutting equipment was the buffered keypunch. This was followed by the key-to-tape machine and later by the key-to-diskette machine, replacing the mechanics of the card punch with electronics and the cardboard with magnetic tape of "floppy disk". It did not take long for disks, multiple work stations, line printers, intelligence, and finally communications to be added (Rhodes, 1980, p. 70).

The facilities offerred by computerized input systems have improved input system efficiency by several orders of magnitude. More and more the word "source" is combined with "data entry" to describe the kind of input that is occurring with current information processing systems. There is a great deal of focus on taking data entry to where data originate to eliminate recapturing it. The trend encompasses placing terminals at user sites instead of having all terminals in a centralized site near the computer. This makes it possible to perform some degree of editing and correction at the point of data entry. Another growing method of source data entry is optical character recognition (OCR), described in Chapter II, where human-readable documents are optically scanned and read into the computer directly, without keying or rekeying. Although optical character readers have shown only a modest gain in usage in recent years, the labour intensive aspect of using keyboards and the rising cost of labour may suggest a wider application of OCR in the future (Lusa, 1979, p. 54).

Today's state-of-the-art data entry systems bear little resemblance to their predecessors. They range from sophisticated

microprocessors which have the ability to edit data at the time of entry to machines which can read handwriting and interpret speech. However, despite changes in technology and trends for the future, many users continue to believe that keypunch is the most cost-effective method of data entry for their particular situations. Nevertheless, the fact that average keystrokes per hour can vary from fewer than 5,000 to over 20,000 indicates the need for a reassessment of management techniques applied in the keypunch environment.

Cost-effective and accurate data entry poses a challenge for the future. Which technology, if any, will emerge as the ultimate method of data entry is still debatable. Data entry applications are extremely diverse and the requirements within each application differ greatly. Most observers concur that it is hard to image any one technique or technology being universally suitable for all data entry applications (Rhodes, 1980, pp. 73-76).

B. Hardware Trends

The following review is largely summarized from United Nations (1980e).

The development of data processing hardware commenced almost 100 years ago with the need to expedite processing of the decennial census conducted by the United States Bureau of the Census. Herman Hollerith was commissioned to build a series of machines that could be used for tabulation of the census results. In contrast to the census of 1880, which took about eight years to complete, the census of 1890 was completed in about two years using Hollerith's new machines.

Modern statistical data processing began with the development in the 1940's of equipment based upon electronic circuitry that was capable of stored program operation. The invention of the stored program allowed for retrieval and execution of program steps at electronic rather than mechanical speeds, and provided the property of self-modification. These advances led to the installation of UNIVAC I, the first commercial computer, at the United States Bureau of the Census for purposes of processing the 1950 Census of Population and Housing.

Data processing hardware is often categorized in terms of the generation to which it belongs. In general, hardware of the first generation was based upon vacuum tube technology, and hardware of the second generation upon transistor technology with discrete components mounted on circuit boards. Successive generation technology has employed medium- and large-scale integration of electronic components using miniaturized and semi-conductor

circuitry based generally upon photolithographic techniques. Because of the multitude of alternatives which have arisen, it becomes less meaningful to categorize equipment in terms of technological generation. Other attributes of the equipment are generally more important, such as capacity, modularity, price, software choices, and ease of operation.

The following discussion looks at trends in five functional components of data processing hardware: processing units, primary memory, secondary memory, output devices, and communications equipment.

1. Processing units

Large-scale integration (LSI) technology has brought about a marked decline in the physical size and cost of central processing units. Advances in LSI fabrication have made it possible to produce entire processing units on a single electronic chip about the size of a human finger tip. This trend in miniaturization is expected to continue at least in the near future, probably yielding 20-30 percent more processing capability each year for the same cost.

A relatively inexpensive outgrowth of the LSI technology is the microprocessor, typically a small processor with a limited data width path and instruction repertoire. The most prevalent on today's market are 16-bit processors with memory management units with capabilities and speeds equivalent to those of large minicomputers of several years ago. The development of 32-bit microprocessors will be the next step. With this development, the typical central processor will exist on one or more small electronic chips and will be relatively inexpensive.

The emergence of inexpensive and relatively powerful microprocessor hardware has fostered a shift away from centralized data processing. The availability of a wide variety of microcomputers and minicomputers, combined with inexpensive primary memory, makes it possible to distribute computing power more effectively at a lower cost. Advantages in efficiency, owing to more simplified software environments and directness of control, far outweigh any lost hardware economies of scale.

2. Primary memory

Primary memory is that memory within a computer system that is most accessible to its processor. Program instructions and data elements of immediate interest are often contained in primary memory for rapid program execution.

While primary memory is frequently referred to as "core" memory, connoting the extensive historical use of magnetic core technology for such memory, the bulk of current primary memory technology relies upon active semi-conductor circuit technology to maintain memory elements. Thus, the same LSI fabrication techniques that advance processor technology also serve to advance primary memory technology. It is claimed that the cost of memory is being halved every three years. In addition to providing memory at lower cost, semi-conductor technology now provides memory products with greater reliability, less power consumption, increased automatic error correction features, and higher speed. These developments also contribute to the decentralization of data processing (see Chapter VI, section A.1).

3. Secondary memory

Secondary memory consists of a variety of storage devices that are used to store data items of less immediate interest to the programs being executed. In comparison to the primary memory, secondary memory is generally more voluminous and cheaper but slower.

The most prevalent current secondary memory devices are magnetic disk and magnetic tape. Disk storage is for the most part accessed randomly, while magnetic tape is accessed sequentially. Other forms of secondary storage consist of magnetic bubble memory, charge-coupled devices, and variations on standard magnetic tape.

Probably the most important recent developments in disk technology have been the introduction of the sealed disk module and non-removable disks. The sealed disk module technology (or Winchester technology) provides a hermetically sealed, and therefore non-contaminated, environment for data transfer and storage. This permits considerably smaller read/write heads with a large potential increase in recording density and module data capacity. For those purposes for which disk modules need not be removed from the computer system, there now exists a variety of non-removable disk products at lower prices and sometimes with better performance for the same capacity than corresponding devices with removable modules.

The expanding microcomputer industry is beginning to exploit Winchester disk technology, and a wide range of new products based on this technology is becoming available. Replication of operating systems previously available at the minicomputer or computer system level are now appearing at the microcomputer level.

Tape technology is also showing some progress. The emergence of a recording mode of 6,250 characters per inch has greatly increased the amount of data that can be stored on a reel of

magnetic tape, although hardware to support this density is still relatively expensive. Further, a new mode of tape recording, known as streaming, had emerged. Streaming permits a rapid transfer of information in large quantities between disk and tape. Such a mode is generally used to backup and restore information on non-removable disks of large capacity.

Magnetic tape technology still offers inexpensive storage of very large volumes of information both for archival purposes and for routine sequential processing tasks, and it does not appear that the medium will be displaced in either of these roles in the near future.

4. Output devices

The range of output devices continues to expand. Largely because of the growth of smaller computer systems, the market in low to medium functionality visual display units (VDU's) and low and medium speed printers has exploded, with low and medium resolution graphic output displays available at reduced cost. At the high end of the spectrum, laser technology is being used to manufacture printers having a significantly higher output rate than that of a mechanical device.

5. Communications

Technical progress in communications hardware is on the whole somewhat slower than that in the computing industry. Nevertheless, progress is being made in the cost-performance characteristics of statistical multiplexers for making more efficient shared use of single communication channels for digital transmission and in the use of faster modulator-demodulator (modem) units at speeds of 1200 baud (120 characters per second) for data transmission. Of more long-range importance is the initial use of optical fibre cable paths, which promises to increase substantially both overall communication capacity and baud rate available to users.

In summary, the prognosis is excellent for increasing effectiveness of computing hardware in statistical data processing activities. Rapid technical progress is increasing the number of alternatives available to the system designer at less prohibitive costs.

C. Software Trends

The term software covers a broad spectrum of programs written to facilitate interaction with computers. These include operating

systems or system software; utilities such as sorts, copy routines, and file maintenance programs; language compilers; and applications software. Although all of these areas are to some degree represented in the following discussion, the trends presented will focus largely on applications software, which includes generalized packages, as this is the area where the majority of programming effort is expended.

1. Quality of software

Software quality is a complex attribute that can be thought of in terms of the dimensions of functionality, engineering, and adaptability. Functionality is the exterior quality - the completeness of the product and the appropriateness of the solution to the user need. Engineering is the interior quality - the reliability and internal performance. Adaptability is concerned with how the system can change to meet new needs and requirements overtime. The combination of these dimensions is complex enough so that no simple quality measure has been developed, and it is not likely that one will be developed (Hetzel and Hetzel, 1977, p. 211). Trends in each of the three dimensions are the best indicators of the progress that has been made over the past years.

Functional quality, i.e. the appropriateness and completeness of software products, is improving; however, gains in functionality have not kept pace with the growth in the complexity of user requirements. Applications in the late 1950's and 1960's tended to involve a single user, with the programmer working very closely with the user. Today, we find multiple users often with conflicting needs and multiple design and programming teams all involved. The new system must fit in with complex existing software structures, and these requirements as well as any time dependent or real time considerations must be addressed. The result is that the problem of fuzzy specifications has steadily worsened. Specifications that are imprecise force testing to be inadequate, and end-user satisfaction suffers. In short, functional quality has not kept pace with system complexity (ibid.).

Better engineering has been reflected in improved program reliability and performance. During the past decade, documentation practices have improved greatly. Major efforts are now made to make programs readable and understandable. New software is now structured or strongly influenced by the principles of structured programming. Overall, software reliability has improved as more emphasis is placed on "fail soft" techniques.

Adaptability has also been greatly improving. The introduction of data base systems, data communications systems, and more abundant and powerful generalized software packages in the

1960's and 1970's have brought about a high degree of independence and greatly facilitated change. Many systems are now generalized enough to handle diverse needs without requiring any recoding.

2. Software development

The major trend affecting software development is the dramatic increase in its cost in contrast to a decline in hardware prices (Cottrell and Ferting, 1978). The challenge confronting the software developer is how to meet the unique processing requirements of the organization without reinventing the wheel with each new appplication or variation of an existing application. This is especially important for organizations engaged in continuing statistical activity with constantly evolving data processing needs. The secret to meeting the challenge lies in moving away from line at a time coding to approaches which increase programmer productivity, reduce need for testing, enhance documentation, and minimize maintenance. These techniques include reusable code systems, application generators, and the use of application software packages where available. All of these purport to increase productivity of application development by:

(a) Minimizing the percent of new software in total software required for a new application.

(b) Extending the life time of a line of code.

(c) Permitting new software to be reusable in other developments.

(d) Reducing the level of skill required for implementation.

(e) Significantly increasing productivity of implementations and quality of resulting products.

(f) Permitting a more direct and unambiguous statement and design of problems to be solved.

(g) Eliminating variability in system design by different individuals.

Reusable code systems involve building a library of precoded complete modules or module skeletons which can be quickly recalled. They alleviate the need to reprogram such things as a new page routine, a standard header, or a two-way match. The utility lies in the fact that such modules require little or no adaption for a specific application.

Application generators allow the programmer to work at a much higher level than is possible with procedural languages such as

COBOL. They shift the emphasis from how something is accomplished to what is accomplished. They can take a variety of forms. In one form the programmer sits down at an interactive terminal and describes in a direct way the particular attributes of his application including output products, input transactions, data relationships, and other external parameters of the application. The "system building machine" has the intelligence to seek fairly complete information from the analyst. Such a generator will then produce highly standardized source code which can be made machine specific with minimum deviation from whatever standard exists for the chosen source language.

Another manifestation of the application generator is the use of a macro language whereby programmers write only macros of executable code and the system subsequently generates source code. One application of this approach claims an average programmer productivity of 4,000 lines of debugged code per month.

Because the system generates the source code, it also achieves a very high level of standardization. This, in turn, eases the maintenance job. In addition, it facilitates system definition and documentation; analysis and design; system test, installation, and production; as well as programming and system maintenance.

A package such as the United States Bureau of the Census' COBOL CONCOR is essentially an application generator to perform data editing and imputation. The user writes CONCOR statements which in turn produce an executable COBOL program which consists of many more lines of code than it was necessary to enter in CONCOR statements. Many of the overhead tasks are transparent to the user.

Application packages allow high productivity because they entail little effort to "produce", and their maintenance is standardized and generally supported by the developer. An analyst often finds these packages to be "user friendly" because the desired output can be obtained with so little effort. However, they are somewhat limited in applicability because the structure of the application is already fully determined by the developer with relatively little room for modification.

In addition to maximizing programmer productivity, it is becoming increasingly necessary to maximize the life of future systems, if only to recover the cost of their development. This implies greater emphasis on software maintainability and hardware-independent implementations possibly sacrificing some "design purity" (Weinberg and Yourdan, 1977). The increase in complexity has had a multiplier effect on the cost of a failure. There is increasing emphasis on the need for reliability beyond simple functional correctness, even at the cost of redundancy.

The contrast between program code efficiency and program maintainability is dramatically illustrated by a study done in 1973 in a controlled experiment of two computer programs of approximately 400 FORTRAN statements each independently prepared to the same specification. One was done by a programmer who was encouraged to maximize code efficiency, and one by a programmer who was encouraged to emphasize simplicity. Ten times as many errors were detected in the "efficient" program over an identical series of 1000 test runs (Swanson, 1976, pp. 592-593).

Many serious gaps remain in the availability of generalized and portable software for household survey data processing. While for data tabulation, and to a lesser extent for some other statistical analysis, there are several packages available, that for data editing and imputation are much less plentiful; packages are almost non-existent for several other applications in survey design and analysis (see Annex I).

There are a number of possible reasons for this. One is the relative infancy of statistical software development. To emphasize this infancy as compared with computer hardware development, it has been said that all available hardware is now either third or fourth generation equipment, whereas statistical packages are now only in their second generation. The first generation of software was typified by the use of independent computation algorithms developed and linked together for one specific machine. Second generation software is easier to use; is more reliable; has greater diversity of statistical capabilities; has routines which are more consistent and consolidated under common controls; and has a higher standard of user documentation (Muller, 1980, p. 159).

Although hardware costs have declined dramatically, software maintenance and development costs have not kept pace. The proliferation of new hardware and new applications has aggravated this software crisis. Even though there may be some applications that are too ill-defined or broad to be suitable for software packaging, portable software packages and interfaces can be developed for many survey applications, thereby providing a standard set of functions for users.

The wide variety of programming languages used to develop software products also has its effect on the lack of standardization of packages. The extensive use of FORTRAN, COBOL, PL/1 and assembler languages does not provide an atmosphere conductive to portability. An alternative might be standardization at the lower level of the software environment for program and package development. The University of California at San Diego has designed a PASCAL system which can be used as a portable software development system for the microcomputer environment. This system, which is comprised of an editor, a file manager, and a debugger, offers much

promise for providing a standard development environment in which interchangeable software products could be produced. PASCAL compilers, unlike other high level language compilers, can work well in as little as 56K bytes of primary storage, enabling them to be installed on almost any computer on the market today (Applebe and Volper, 1979, p. 117-118).

In lieu of standardized PASCAL compilers for most major vendors world-wide, the software developer's increased use of independent COBOL will go a long way to making packages portable and standardized. Independent COBOL relies on the internal use of pseudocodes (non-specific device references) within each particular programmed module for which a standard operating system interface (resolution of non-specifc code) is supplied by each COBOL vendor (Taylor, 1980, p. 31).

Another major area which has had a serious effect on the perceived advantages of software packages is that of maintenance costs. Too often the potential user steers away from the acquisition of a product, which may be very useful, because of a fear of heavy maintenance costs.

3. Development of integrated systems

Several large or advanced national statistical offices have commenced development of programs for a unified, integrated statistical software system. As this development continues, smaller statistical offices may perhaps also benefit from the technologies that arise. Data base management systems, data tabulation systems, and data presentation systems are all useful in their own right, but an integration of these systems is essential to increase user-orientation.

In these systems, data are entered through the data base management system (DBMS) into the data base where they can be edited and imputed. The data tabulation system accesses the data through the DBMS, and stores the resulting tabulations back into the data base. The tabulation results can, of course, be immediately printed as tables for examination, and can be used as input to the data presentation system for the preparation of charts and statistical maps. The common user interface allows users of the total system to deal with single, uniform sets of concepts, terminology, and procedures for carrying out data tabulation and presentation. Examples are provided by Statistics Canada and United States Bureau of Labor Statistics.

Statistics Canada has two partially integrated systems. The first one, which produces working tables, utilizes RAPID, the relational data base system, with STATPAK, the table generation

system that works with the data base. The second system does photocomposition using the table generator system CASPER with some custom-coding to interface with videocomposition equipment owned by a private contract firm.

The United States Bureau of Labor Statistics (BLS) created a system which uses the network data base management system TOTAL, with BLS's own generalized tabulation package TPL. Photocomposition is done using PCL, the print control language within TPL. The resultant output is phototypeset using the United States Government Printing Office Linotron. BLS is working toward a completely integrated system.

It is interesting to note that the highest degrees of integration in existing systems are for small, limited-purpose systems. This is so because it is difficult to integrate existing systems which were not initially designed to be integrated, and it is too expensive to develop new systems. Systems integration is also hampered by portability and standards problems (Alsbrook and Foley, 1977, pp. 63-75).

If integrated statistical software systems having a common user interface are going to become the standard for the future in statistical offices, then software developers must recognize three different levels of user programming: the statistical language, the algorithmic language, and the interface language.

The statistical language is what the user sees. It should have the potential, at least, for analysis in an interactive mode, with immediate feedback and graphical output. It should emphasize simplicity, relieving the user of inessential details, providing security against errors, and allowing and encouraging insightful data analysis, with informative feedback and few restrictions.

The underlying support for any extensive system should be a set of algorithms. These will provide the numerical calculations, such as solving least-squares problems or generating pseudo-random numbers. The algorithms should be logically correct and well tested. The methods used should be reliable and reasonably efficient. If algorithms are to be shared, they must be reasonably portable.

The interface is the software which links the statistical language with the underlying algorithms. The interface must contain the code to interpret the user requests. It is important that the system designer's time be well used by making the interface writing as easy as possible (Chambers, 1979, p. 100).

It would be ideal if advanced statistical organizations were willing to make the large initial investment to develop statistical systems that would offer the range of statistical operations needed to process complex sample surveys. There is, however, general pessimism in the statistical software field regarding the possibility of a generalized system of programs with transparent interfaces that could address all the processing needs of such surveys. As noted above, a few examples exist of large government statistical offices endeavoring to assemble integrated statistical systems for processing censuses and surveys, but typical national statistical offices are unlikely to benefit directly from these efforts because the systems were not designed with portability in mind.

In recent years, considerable attention has been given to the quality of software in national statistical offices by the Conference of European Statisticians (CES) of the United Nations Statistical Commission and the Economic Commission for Europe (ECE). A Working Party on Electronic Data Processing of the CES has for several years prepared reports and held meetings with representatives from the national statistical offices of all member countries on the various aspects of data processing. Working with the Computer Research Center in Czechoslovakia, considerable attention has been given to developing an Integrated Statistical Information System (ISIS), parts of which are already in use in the statistical offices of other countries. A group within the Working Party prepared a report recommending that the national offices prepare a joint statement on the specific characteristics of statistical data bases, and that model software for statistical DBMS systems be developed. The CES is proposing to establish a clearing-house where national statistical offices would deposit copies of generalized programs, which would be transmitted to other national offices on request (Alsbrooks and Foley, 1977, pp. 63-65).

4. Standards for software development

In conclusion, it should be instructive to summarize the considered opinion of software evaluators as to the standards which should be followed in software development, whether single-purpose routines or integrated statistical systems. While these standards are normative, they also indicate the trend in software development insofar as the producers increasingly try to meet those standards.

(a) Language should become more user-oriented, with understandable syntax for describing the necessary statistical tasks without the use of computational or procedural details.

(b) Future packages should be able to handle unrestricted types and quantities of different inputs. Data should

be identified according to their source, quality, editing conditions, and timeliness in order to facilitate future retrieval.

(c) Provision should be made for simultaneously handling multiple versions of the data, such as historical and current data.

(d) Controls and monitoring information should be provided for handling in a consistent manner missing data and for indicating whether the data are compatible with the assumptions required of the algorithms that have been used in the package.

(e) The user should be able to choose among alternative algorithms for analysis.

(f) It is reasonable to expect improved quality and flexibility in preparing reports that are more attractive and more readable than those currently prepared on impact printers. It would be desirable for a package to have a report control language that would make the specification of reports much simpler than at present with regard to format, content, layout and particular hardware devices to be used.

(g) Extensibility of a package is often desirable to permit the user to augment its existing routines to handle the particular data or analysis.

(h) More effort should be made to achieve true portability, encompassing the data, the programs, the test cases, and the documentation.

(i) It would be desirable to see some means for the user to evaluate the performance of a package.

(j) Maintainability should be emphasized since an increasing variety of equipment, including new and distributed hardware, is involved. Economies of scale can be realized from centralized maintenance.

(k) It would be desirable to have testing facilities included in a package to give the user ways of validating it; that is, special routines or a test made that would assist one in testing the package.

(l) More attention should be given to documentation, including development of performance documentation to aid those who want to use, modify, or maintain the package.

(m) It would be desirable to have packages function in multiple modes, including batch, interactive, diagnostic test, and tutorial modes (Muller, 1980, pp. 161-163).

V. EXISTING DATA PROCESSING SITUATION IN NATIONAL STATISTICAL OFFICES OF DEVELOPING COUNTRIES

This chapter must be prefaced by stating that there is no typical country or stereotype to describe the data processing situation in the developing world. The material presented is intended to give the reader a general feeling for several aspects of data processing in developing countries. In some cases these ideas can be substantiated by providing the number of countries to which they apply.

A. Data Entry

Data entry has traditionally been a bottle-neck in the processing cycle. Large efforts, such as national censuses, have been plagued by the scarcity of equipment to accomplish data entry in a timely manner. The data entry load for household sample surveys can, of course, be expected to be smaller than full scale censuses, though still requiring very considerable time and resources for continuing programmes.

Most countries have begun to shift away from using conventional keypunches toward keying to a magnetic medium such as a diskette, cassette, or central disk. This is in keeping with the general trend in data entry although the shift in developing countries has been much more recent. The high cost of punch cards and the fact that they are not reusable was one contributing factor to this change. The ease with which changes could be made to the keyed data and increased productivity offerred by the newer equipment also influenced the decision. However, optical mark reader (OMR) equipment is rarely used.

The abandonment of the keypunch has brought about a situation where it is difficult to obtain service and replacement parts for the older machines. In some cases, keypunch machines are being "cannibalized" to provide parts to repair others.

The newer data entry equipment often has the ability to be programmed for some degree of editing. However, most countries are either unable to utilize these added capabilities or choose not to take advantage of them, using these machines as though they were simple keypunch machines.

Maintenance is very much a problem. It is rare to find all data entry equipment functioning well at any point in time. Spare parts must often be ordered, leaving the malfunctioning equipment idle until they arrive. Some of the maintenance problems can be attributed to a general lack of trained service technicians.

The newer equipment sometimes poses a problem which did not exist when the input medium was the punched cared, namely, difficulties in the transfer of data from the entry medium into the computer. There are two possible conversion problems:

(a) The machine has no peripheral device to read the diskette or cassette, in which case the data must pass through a converter which puts them on magnetic tape.

(b) The data entry equipment records in EBCDIC and the computer operates in ASCII, EBCDIC and ASCII being two schemes for representing data on magnetic media. In this case, a program must convert each keyed entry to the appropriate code so that it can be understood by the machine.

Production speed and quality are variable. The average speed can range from 5,000 to 15,000 key strokes per hour. This indicates the importance of proper management. Also, the speed and the quality of work produced are greatly affected by the training provided and motivation to do a good job.

B. Hardware

1. Access to computer equipment

Computer equipment used by national statistical offices around the world can be described as a vast range of second, third and fourth generation equipment. It is quite accurate to say that today almost all national statistical offices have access (whether in-house or at some other location) to electronic data processing equipment for at least some part of their statistical processing workload.

Over the recent past the dramatic drop in the cost of mainframe hardware coupled with the large increase in hardware vendors operating internationally has made it possible for many statistical organizations to upgrade their equipment from second to third or from third to fourth generation equipment, or to acquire computers for the very first time. Reduction of the "million dollar plus" price tag usually associated with large-scale equipment down to a mere fraction of that cost for medium-sized fourth generation equipment (which outperforms large-scale third generation equipment) has opened the way for many smaller statistical offices to be able to afford to procure their own computers.

In the African region a 1978-79 study administered by the United Nations Economic Commission for Africa (United Nations,

1980d) reported that of the 17 countries which responded to a questionnaire, only six countries had computer installations located at the central statistical offices. Three of these countries needed supplementation of their processing capabilities by other government agencies. The two common alternatives to a national statistical office doing statistical data processing in-house seem to be using the computers at the ministries of finance or at national data processing centres. In a few cases, the statistics offices used machines of other government agencies or shipped their data abroad for processing.

The ECA report leaves the impression that there is definite advantage for a statistical office to have its own computer equipment even if it cannot utilize the full capacity of the machine. All statistical offices in the study which had to go to other agencies for computer services experienced serious delays in accomplishing their work as a result of having to accept lower priorities to the host installation's own applications. On the other hand, statistical offices having their own equipment could offer unused resources of the machine to other government applications, such as payroll and government accounts, while at the same time keeping a high priority for their own work (ibid., p. 12).

The recent expansion of overseas operations by major hardware vendors has led to a greater vendor competition in many countries, probably resulting in improved service and support to the user organizations. A study done by the United States Bureau of the Census in March 1980 illustrated IBM's shrinking hold on the international computer market, especially in government-run statistical data processing centres. Information available from 98 countries covering 170 computer installations which were either government-run or government-used showed that only 63 percent had IBM products. Ten years earlier IBM virtually controlled the overseas market, and as recently as five years earlier IBM still maintained a 90 percent share of that market. ICL equipment ran second to IBM, being in place at 14 percent of those installations. Following ICL were NCR and Honeywell, with 7 percent and 5 percent respectively. A few UNIVAC, WANG, FACOM, NEC, CDC, and Burroughs computers were also found. ICL computers showed predominance in Africa and East Asia and did not appear at all in Latin America. A few ICL machines were found in Western Asia and the Caribbean.

Although the 1978-79 ECA report on African Statistical Data Processing included responses from all organizations having computer facilities (not restricted to national statistical offices), the trends of predominance by certain hardware vendors in the region can be applied specifically to the national statistical offices. The report showed IBM products installed at 44 percent of the sites with ICL equipment existing at 23 percent of the sites. Honeywell Bull equipment was the third most prevalent equipment found, in place in 14 percent of the installations. Burroughs, NCR, and Hewlett-Packard held less significant portions of the market in the region.

2. Computer capacity

A very general measuring stick of a computer's size or capacity is the size of the primary storage available. Primary storage is particularly important when determining which software packages can be installed on a machine. Only as more and more virtual storage machines become available in developing countries will the size of primary storage diminish in importance. The United States Bureau of the Census study mentioned above showed a vast range of memory sizes, from 8K bytes to 3 megabytes. The average memory size was an impressive 333K bytes which would indicate at first glance that most installations have an abundance of storage, capable of processing complex surveys and capable of hosting complex software packages. Closer inspection of storage capacities at the regional level indicated a significant difference among regions. Western Asia and East Asia had the largest machines, with an average of 758K bytes and 586K bytes of storage respectively. Latin America was in the middle, with an average of 325K bytes of storage while Africa, South Asia, and the Caribbean were at the low end, with 174K bytes, 78K bytes, and 78K bytes respectively.

Core storage capacity was reported by 129 installations in the ECA study and approximately 56 percent of African installations were of less than 100K bytes with the most frequent primary storage capacity ranging from 32K bytes to 128K bytes. There is still a tendency for small to medium core central processing units, but there appears to be a move towards upgrading some of these units. All sites that reported new machine acquisition plans for 1980 called for central processors having primary storage capacities well in excess of 128K bytes.

In terms of peripheral devices, most countries have magnetic disk and tape units available, which facilitate processing and storage of large or complex surveys. On some smaller machines that have a restricted amount of on-line disk storage, it would be necessary to store most final output files and some intermediate files on tape rather than on disk. This, of course, slows down the processing rate and adds to the complexity of operational control but, nevertheless, does allow most types of surveys to be processed successfully in principle.

C. Software

If national statistical office computer centres in developing countries are lacking in any aspect of current data processing technologies, it would be in the area of acquired software. Most organizations have high level compiler languages on their machines which aid them in developing complete custom systems to process their surveys. Unfortunately, many of these installations have not

taken advantage of available general statistical software products which could significantly lessen the burden on their already overcommitted programming staffs to prepare systems to process surveys on a timely basis. The limited use of software packages in developing countries cannot be attributed to the absence of COBOL and FORTRAN compilers which are the host languages for most packages. Most installations now offer one or both of these languages.

In terms of usage of compiler languages, the ECA report states that 82 percent of the establishments participating in the African study used COBOL quite extensively, and of those using COBOL, all but six also used FORTRAN. RPG was found to be used at one-third of the sites and ALGOL, PL/1, ASSEMBLY, PLAN, BASIC, NEAT, and AUTOCODER were used much less frequently at a few other sites.

According to a survey of national statistical computer installations undertaken by the Economic and Social Commission for Asia and the Pacific (ESCAP), all but one of the 17 countries in the region used FORTRAN and all but two used COBOL. RPG-II was reported used in nine countries while PL/1 in four (United Nations, 1978b).

In the area of generalized statistical software packages used by national statistical offices in developing countries, the overwhelming majority of installations that use any kind of packaged software use editing and tabulation systems developed by the United Nations Statistical Office and the International Statistical Programs Center of the United States Bureau of the Census. This may be due, in part, to the fact that both the UNSO and ISPC design their packages specifically for use in developing countries and they do not hold proprietary rights to the software (they do not charge anyone for its use).

The United Nations Statistical Office (UNSO) over the past several years has been actively involved in the delivery of the edit and tabulation packages UNEDIT and XTALLY to many developing country statistical offices (see Annex I for a description of the various software packages). During 1978-80, the UNSO delivered UNEDIT to 21 developing countries and XTALLY to 22 countries. In some cases the software was installed by local staff following written directions, but in most cases, it was installed and demonstrated either by staff of the UNSO or by a United Nations regional data processing adviser familiar with the details of the programs and computer operating systems. In all cases, the installation sites were computer centres at national statistical offices or at different government agencies at which census processing takes place. These packages have been installed on a range of equipment from IBM S/3 Model 10's to IBM S/370 Model 135's. As of 1981, the UNSO had a log of outstanding requests for UNEDIT and XTALLY from 31 other developing countries and it was planned that the software would be delivered as soon as

possible to 24 of those countries. The remaining countries either had machines with insufficient primary storage or no RPG-II compiler, which prohibited the installation of the software (United Nations, 1980e).

The International Statistical Programs Center (ISPC) of the United States Bureau of the Census has been delivering software to developing country statistical offices for more than ten years. Through the auspices of the United States Agency for International Development (AID), ISPC developed the general tabulation systems CENTS (IBM Assembler) and COCENTS (COBOL) as part of the 1970 World Census Programme and installed these systems either directly or through third parties (such as UNSO) in many countries. The latest United States Bureau of the Census survey (1980) shows CENTS being used in 47 installations and COCENTS being used in 65 installations world-wide. The distribution of this software across regions was quite even, as at least one of the two packages was in place at eleven sites in the Western Asia, eight sites in South Asia, nineteen in East Asia, thirty one in Latin America, four in the Caribbean, and thirty five in Africa. Recently ISPC completed the latest version of its comprehensive edit and imputation system, COBOL CONCOR, and a programme of distribution was commenced by AID through a private contractor, the NTS Research Corporation of Durham, North Carolina.

When questioned about the use of editing software in the ESCAP survey mentioned earlier, seven countries indicated their preference for the use of CONCOR while one indicated probable use of UNEDIT. The remaining countries were either undecided or stated their desire to use specific custom programs or unspecified generalized packages. In the case of tabulation software for survey processing, ten countries indicated their preference for CENTS or COCENTS, while other countries individually specified XTALLY, FILAN, ICL Survey Analysis/FIND-2, MACR-PACKAGE, FTL6, TPL, or SPSS.

In the ECA report regarding specific software packages, there seemed to be a rather low level of use of such packages in the region. The most prevalent systems used were CENTS, COCENTS and XTALLY for table generation. To a lesser extent, SPSS or a modification of it was used as well as FIND-2, a package for multiple file inquiry.

D. Typical Problems

1. Staffing

The one statement that can be made about developing countries, almost without exception, is that data processing personnel are grossly underpaid in comparison to what persons in

similar positions receive in the private sector in their countries. For this reason, the national statistical office tends to serve as a "training ground" from which employees move on to more lucrative positions in the private sector.

The capabilities of the available professional staff cover a broad range. Most are not university graduates, although it is common to find analysts and programmers studying at the local university. This training tends to be rather formal, with insufficient practical experience to reinforce what is learned. Some specialized training may be provided by the supplier of computer hardware and software. The best individuals have often acquired their expertise by on-the-job experience and trial-and-error, and are virtually indispenable because of the versatility they have acquired.

Most employees could profit by additional training to supplement the limited training and experience they possess. Unfortunately, some agencies view training as a dangerous thing since it can encourage the employee to leave the government organization. International training is almost guaranteed to generate job offers upon the employee's return.

Technical manuals are expensive and difficult to obtain. It is rare that an office has one complete set of current manuals. For this reason, employees carefully guard their manuals and always like to obtain additional manuals.

There is generally a low level of motivation among data processing personnel. This is brought about by a combination of factors:

(a) The relatively low salaries are a constant reminder that they are not being paid at the same rate as their counterparts in the private sector.

(b) Communication with subject-matter personnel is generally inadequate and the data processors are not consulted about matters which affect their work. They often feel that they are working in isolation because of this lack of involvement.

(c) There is little recognition by management of the work done by the data processing staff because of a lack of understanding of the substance of the work. Their success is judged solely on their ability to produce the needed products, but how this is accomplished is largely ignored.

These factors do not support the development of highly motivated and innovative staff, although such individuals do exist in many statistical offices.

As noted above, the personnel problem of trying to hire and retain good people is probably the most critical. It is often very difficult to get permission to hire new staff even though there is a severe shortage of data processors. Increasing salaries of existing staff may be an impossibility because of the need to retain parity across various government agencies. As a result, most national statistical offices suffer from a high rate of turnover in personnel and the inability to attract qualified professionals with previous experience. Offering salaries competitive to those found in private industry will go a long way toward cutting down personnel turnover. However, it is often necessary and more important to look for non-monetary solutions to the turnover problem such as increasing motivation through greater responsibility and training, reorganizing to alleviate personnel incompatibilities, improving the work environment, giving the employees more participation in making decisions, and giving recognition for achievement at all levels.

2. Access to computer

Securing access to a computer and getting good turnaround are critical to the timely completion of any data processing project. National statistical offices which do not have a computer for their exclusive use often experience difficulty in having a machine available for their use and in getting results back quickly. Frequently, this problem can be traced to poor management of the computer facility. The machine may be quite adequate in terms of its capacity, but problems in priority setting, improper operation, and lack of concern for the user may prove to be great frustrations. It is this lack of user control that makes every agency want its own computer.

3. Lack of vendor support

Lack of responsiveness on the part of the vendor, i.e. supplier of computer hardware and software, can manifest itself in several areas. Maintenance is probably the most critical area, since malfunctioning equipment can affect so many people. Vendors who do not have strong competition often do not feel compelled to answer service calls promptly, to adequately train their repair persons, to expedite acquiring replacement parts, or to provide preventive maintenance. Vendors are frequently unprepared to answer technical questions on their hardware or software or to look into problems in these areas.

Sometimes there is very little choice in selected equipment because only one or two vendors may be represented. Unless the national statistical office wants itself to take on maintenance responsibilities, it must select from the local market which may not offer many alternatives. In addition, some companies market only a subset of their complete product line in developing countries in order to cut down on maintenance and training requirements.

4. Unstable power supply and inadequate facilities

Problems with instability in electrical current can plague the users of the machine. The fluctuation may be severe enough to completely disable the computer or may cause unpredictable damage if a drop in power is not detected. Frequent electrical failure is not only frustrating to the users but may cause severe damage to the machine. Installation of equipment to deal with electrical failure or fluctuation may be necessary if the problem is severe.

Inadequate facilities may be a problem for the equipment and the staff. There are certain environmental requirements for proper maintenance of the equipment and its related supplies. Improper control of temperature, humidity, and air quality are often responsible for machine failure and damaged supplies. Inadequate ventilation and lighting and the general lack of a good working environment contribute to dissatisfaction and low productivity among staff members.

5. Lack of realistic planning

Data processors in developing countries often lack the experience needed to make realistic schedules. There is a tendency to plan for the best case, not the worst. A turnover in staff and uncertainty in accessing the computer make accurate planning difficult, if not impossible. Moreover, realistic planning may sometimes be contrary to cultural practices, for example, where everyone knows to increase estimates by 300 percent, but only the optimistic schedule is presented in writing.

The prevalence of outdated approaches to data processing and a reluctance to accept change may be a problem. Editing and correction techniques are often not understood or are not implemented correctly. This implies the need for an orientation or reorientation to good survey processing techniques before such a staff could be able to successfully process an ongoing household survey programme.

These problems taken collectively portray a dim picture of work in the developing world. However, it should be emphasized that a particular country may experience only a few or even none of these

problems, and they are certainly not unique to the developing world. Just as there are typical problems, there are also successes. In fact, it is often the recurring problems which spawn innovative approaches to their solution. For example, equipment failure has prompted the manufacture of temporary spare parts and the swapping back and forth of critical components. Outdated operating systems are often kept in reserve, for example, to revert to tape operation when disks are damaged or to compensate for loss of other components of the system. The high turnover of personnel has prompted the need for good documentation. Many national statistical offices enforce rigid standards for documenting ongoing systems in order to cope with the change in personnel. Good supervision and motivational techniques have in many situations resulted in high production rates and excellent quality in the data entry activities. Many national statistical offices pride themselves on continuing success in the area of data entry.

The most obvious proof of success is the fact that work continues to be done and dedicated staff members remain despite a low operating budget. It is easy to get discouraged by the lack of importance given to the generation of statistics, but most countries continue to cope with the problem and persevere.

VI. BUILDING DATA PROCESSING CAPABILITY

This chapter discusses the various considerations involved in the choice of appropriate strategy so as to ensure timely processing of the data generated by continuing survey activity, and at the same time to create or enhance data processing capability.

The building up of data processing capability requires, among other things:

(a) Proper organization of data processing facilities, ensuring efficiency and, even more importantly, effectiveness.

(b) Determination of the appropriate scope and strategy of, and priorities within, the EDP task, taking into account both the user requirements and the availability of means to meet them adequately.

(c) Appropriate choice of hardware equipment in the upgrading of computer facilities as necessary.

(d) Acquisition of packaged software suitable for survey data analysis, and the in-house development of software where required.

(e) Recruitment and retention of good quality staff, and above all the provision of on-the-job as well as formal training at all levels.

A. Organization of Data Processing Facilities

In establishing or strengthening data processing capability one of the fundamental decisions to be made concerns where and with what means data processing will be carried out and how the facilities will be organized. These issues immediately suggest the question of centralization versuses decentralization which manifests itself at two levels:

(a) Whether a central or national centre is to be used, as opposed to in-house computer facilities within the statistical agencies.

(b) Whether the various tasks should be carried out on a single computer or in one place, as opposed to the distribution of these activities to a number of sites or locations.

A related question is whether the statistical agency can carry out its own data processing, or whether it will be necessary to contract the task to an outside body.

While some of the issues involved may be beyond the influence of the national statistical agency, the undertaking of a continuing programme of surveys may provide an opportunity for, as well as necessitate, a reassessment of the way the computer facilities are organized.

1. Centralized processing versus in-house facilities

Freestanding, or self-contained, computers were the norm around 1960. The idea that predominated at the time was to have isolated data processing units to solve local problems. However, at certain activity levels, they were more expensive than large-scale, centralized computers, and the information flow to higher levels was slow or otherwise deficient. Overtime, these shortcomings in freestanding computers, along with advances in computer technology, fueled a trend toward centralized processing; that is, the use of powerful central processing units with large, rapidly accessible files, to and from which all information flowed (Kaufman, 1978, p. 9). This trend was supported by "Grosch's Law", which refers to an empirical finding, valid for the first two decades of computer use, that the raw computing power of hardware was proportional to the

square to its cost. This relationship supported the centralization of processing on the basis of economy of scale in hardware purchase for all job mixes requiring substantial computer power.

However, since the mid-sixties, a number of developments have tended to reverse this relationship. The two important contributing factors have been:

(a) The changing capital/labour ratio, which has seen computing hardware decrease in cost while human resources became more expensive.

(b) An increasing emphasis on effectiveness of the computer component over merely the cost of computations.

Moreover, the user of the centralized system was paying a price both in terms of increased costs of executive level software and in terms of complexity of access and use (Sadowsky, 1977, p. 20). As Richard Canning, one of the respected commentators in the area of data processing, has observed: "One of the lessons of the first two decades of computer use has been: 'Big Systems often mean Big Troubles'".

In consequence, a widespread desire has arisen among data processors for an alternative to centralized processing, but not a return to the freestanding mode. The compromise has been described as "distributed processing"; that is, the deployment of computerized data processing functions where they can be performed most effectively, at costs usually less than those of other options, through the electronic interconnection of computers and terminals, arranged in a network adapted to the user's characteristics (Kaufman, 1978, p. 10). Several technological advances have facilitated distributed processing: the superchip, which has enabled microprocessing; advanced teleprocessing; new era software, including widely applicable generalized systems; and a wide variety of peripheral hardware.

There are economic and non-economic advantages to the distributed processing approach. There is now a widespread feeling that, in many cases, the high powered small machine operating in a dedicated manner produces a cheaper unit of work than the large machine encumbered with its extensive overhead commitment. If use of a distributed system significantly increases the value of the information received, the net result may be highly cost-effective processing. Distributed processing provides the best means for local management to focus directly on controllable conditions, enabling higher management to concentrate on, evaluate, and resolve larger issues. Whereas centralized processing has often given rise to unhappy situations because personnel believe they have been deprived of control, distributed processing can provide a processing environment which has most of the advantages of stand-along autonomy (ibid., p. 11).

However, caution must be exercised in moving away from centralization to ensure maintenance of standards and compatibility among installations and to avoid technical isolation.

In addressing the first centralization issue as it applies to processing a continuing household survey in light of the above discussion of trends in this area, it would seem that in-house processing is preferred over processing at a central computer centre outside the statistical agency. This decision, of course, must take into account local conditions. There is quite a difference between moving a data processing operation from an existing central site to a smaller in-house site which is already operational, and setting up data processing capability where none existed previously. The proper infrastructure must be in place in order to support an independent system.

If a small computer is to be used, there are two features which would greatly enhance its usefulness. First, it would be advantageous to be able to hook it on-line to a larger, perhaps central, computer for applications needing more core or software not available on the stand-alone machine. Second, it would be desirable to produce output which could be easily transferred to other computers on magnetic tape, disk, or diskette. These two features combine to increase the versatility of the small machine.

2. Centralization versus distribution of tasks

The second centralization issue looks at where various processing tasks are to be located. In one sense, the ideal situation would be to edit the data as they are collected so as to be able to take advantage of respondent presence. However, the idea of a fleet of interviewers with microcomputers in backpacks utilizing a digitizer for data entry is far from practical for cost and control reasons - the essential field editing has to be done manually. The next alternative would be to conduct data entry and possibly editing in regional offices. Unless this entails going back to the field to resolve errors, it simply multiplies the control problem without adding any significant advantages. Generally it would seem preferable to carry out all phases of data processing at one site or at as few sites as possible. The fewer separate locations, the easier it is to assure uniformity of operational procedures, maintain control over the flow of work, and establish effective communication. Furthermore, qualified professional, managerial, and supervisory personnel are almost always in short supply, and the fewer locations that have to be staffed the easier it will be to obtain the full complement of personnel needed at each site to ensure a successful operation (United States Bureau of the Census, 1979, p. 116).

This does not imply the full integration of all processing activities at that site, however. For example, there are several good reasons to support off-line data entry, including:

(a) A failure of the computer does not disable the data entry operation.

(b) The data entry process does not interfere with other processing.

There are strong arguments to support centralization of systems development and programming:

(a) The very nature of an integrated system of programs implies the necessity of close communication.

(b) Control over programming and documentation standards can be better maintained in a central location.

(c) Decisions affecting the entire system will be channelled through a central person or group of persons.

The question of centralization versus decentralization or distributed processing should be addressed in terms of efficiency and effectiveness: each situation merits individual consideration in seeking the best approach.

3. Contracting out

The objectives of building up enduring capability and undertaking a continuing programme of household surveys are clearly incompatible with the statistical agency contracting out the data processing task to an outside body. However, in extreme cases, where data processing resources are nonexistent, scarce, or are fully committed to other work, contracting out may prove to be the only way in the short run to get the job done. At the most, this can be a solution for one or more particular surveys but never for the survey programme.

Even for particular surveys, contracting out can involve serious pitfalls. Initially, the idea may sound appealing because it appears to transfer the responsibility for a difficult task to another entity. However, the fallacy in this thinking is that the ultimate responsibility can never be transferred. Furthermore, there can be a number of other problems:

(a) The contractor usually works according to a written agreement. Such an agreement must be carefully thought out in order to cover all details of the relationship. However, there is a vicious circle: the lower the statistical agency's own capability, the more difficult it would be for it to establish and supervise a satisfactory contractual agreement.

(b) If the contract is awarded by the lowest bid, it may not be possible to secure the best person or group for the job.

(c) The contractor often has little or no orientation toward the subject-matter of the survey, and must be given a basic understanding before starting work.

(d) Generally, contractors are in business to make a profit; therefore, they may not provide the most cost-effective means for accomplishing the processing task.

(e) Maintaining communication and monitoring of the contractor's work can often become very time-consuming.

(f) Above all, contracting out provides little opportunity for institutionalization of data processing.

4. Renting versus buying

A similar question relates to renting of computer facilities versus outright purchase. For ongoing activities, the second alternative is, of course, preferable: renting, even over a limited period, can often turn out to be more expensive, and also does not provide protection against price increases. However, renting can sometimes be justified as a temporary measure:

(a) When the resources required for outright purchase are not immediately at hand.

(b) When it is necessary and preferable to wait for the availability of a configuration more suitable for the tasks.

B. Choice of Data Processing Strategy

1. Variation in country needs and circumstances

In developing an approach to processing the data from an ongoing household survey programme, a country is wise to study the

methods and procedures used by other countries in processing similar survey data. However, it will quickly become apparent that no two countries are in exactly the same situation or follow the same approach. It is important to remember that even though a method or procedure is successful in one country, it may not be appropriate for use by another country. In studying alternatives, countries must seek those which make the best use of their resources and meet their needs.

Methods and procedures can be categorized in two groups: those which require increased time, money, personnel resources and facilities; and those which do not require substantial additional inputs. Examples of the first group would be buying a new computer, matching survey data to data from administrative records, or writing generalized software for editing. Examples of the second group are designing precoded questionnaires, implementing operational and quality control procedures, and upholding good management practices in the computer centre.

In evaluating and adapting others' practices, a country should first of all focus on those methods and procedures which can increase efficiency without requiring many additional resources. In considering alternatives requiring substantial additional resources, the various possibilities should be evaluated from many points of view. It is not enough to say "country X uses this approach so we must do the same" or "this is what we must do to stay abreast with modern technology".

The following profiles of seven household surveys from four countries serve to illustrate the wide range of variation that exists between survey requirements, available facilities and practices.

Survey A

87,000 addresses monthly

Computerized sample selection with 1/8 rotation each month

Core of labour force questions with rotating modules

No linkage between survey rounds

Independent verification of industry and occupation coding

Film Optical Sensing Devices for Input to Computers (FOSDIC) data entry

Own computer with 16MB of core memory

Customized programs for editing, tabulation, and estimation of variance

Automated correction of edit rejects

Processing cycle of eleven days

Survey B

30,000 households monthly

Manual sample selection with 1/8 rotation each month

Core of labour force questions with rotating modules

Key-to-disk data entry using 160 machines

Own computer with 32K words of core memory

Own packages for editing, tabulation, and analysis

Manual correction of edit rejects

SPSS for estimation of variance

Survey C

14,560 households yearly

Manual sample selection with 1/3 rotation each year

Two questionnaires: household and individual with rotating modules

No linkage between survey rounds

Keypunch data entry with 100 percent verification

Own computer with 256K words of core memory

FILAN package for editing and tabulation

Manual correction of edit rejects

Customized program for estimation of variance

Processing cycle of ten months

Survey D

26,820 households yearly

Manual sample selection with rotation every two-three years

Core questionnaire with rotating modules

No linkage between survey rounds

Key-to-tape data entry with sample verification

Own computer with 96K of core memory

Customized programs for editing, tabulation, analysis, and estimation of variance

COCENTS package for tabulation also

In the process of establishing a data bank

Processing time of one year

Survey E

22,000 households monthly

Computerized sample selection with partial rotation

Labour force data collected using same questionnaire each month

No linkage between survey rounds

Key-to-diskette data entry

Own computer with 2MB of core memory

Customized programs for editing, tabulation, estimation of variance, and analysis

Manual correction of edit rejects

Processing cycle of eighteen days

Survey F

55,000 households monthly

Computerized sample selection with 1/6 rotation each month

Three questionnaires used; core of labour force data; supplementary surveys questionnaire varies monthly

Data linked for household for six-month period for which it is in sample

Key-to-disk data entry with 112 terminals

Own computer with 8MB of core memory

Customized programs for editing, tabulation, and estimation of variance

Combination of manual and automated correction of edit rejects

Processing cycle of nineteen days

Survey G

1,000 households monthly

Manual sample selection with no rotation

Same questionnaire for each round containing household member characteristics, housing characteristics, expenditures, income, and other socio-economic information

No linkage between survey rounds

Data are transcribed and then keyed to punch cards using 15 machines

Own computer with 2MB of core memory

Own package for editing

Customized programs for tabulation

SPSS and SAS for estimation of variance

Manual correction of edit rejects

Processing cycle of eight months

2. Major factors determining data processing strategy

(a) Scope of work

The complexity, size and frequency of the surveys is, of course, the primary factor which, in view of the currently or

potentially available facilities, determines the appropriate data processing strategy.

However, as noted in Chapter III, this is a two way process: the scope of data collection cannot be determined independently of the possibilities of its timely processing. For example, in a multi-round survey, even relatively minor changes in the questionnaire between rounds can substantially increase the data processing load by requiring extensive modifications in documents and programs. It may be necessary to temper substantive consideration for such change and make some apparent sacrifices in order to meet the requirement for timely data on a continuing basis.

Once the scope of the survey programme is determined, it may still be necessary to make choices and determine priorities within the data processing operation, insofar as all that may be ideally desired cannot be accomplished. For example, first priority may be given to the initial tabulation of all the data collected rather than to microlevel linkage of some data. Similarly, it may be possible only to perform sample verification of data entry or to opt for automatic correction of edit rejects, if there is no time to employ 100 percent verification or to manually review all edit rejects.

(b) Available budget, equipment and software

Restrictions in the existing equipment or limitations in its expansion are of fundamental importance in the choice of appropriate methods and procedures. A computer, by virtue of its core memory size or available compilers, may not be able to support certain software packages. The speed of the machine may dictate the time needed to process the data. If new or additional equipment can be acquired, this may pose no problem; however, if this is not possible, it is important to understand the restrictions of the existing equipment and how they affect the desired approach.

Next, computer equipment requires considerable vendor support. It is important to investigate the availability of responsive local support or proper training facilities for the staff to provide that support. A malfunctioning piece of equipment, regardless of its supposed capabilities, is worthless. The provision for support is essential to the procurement of any equipment.

Certain techniques and software will demand technical assistance and training in order that a country might fully utilize its capabilities. The requisite assistance and training should be assured before these techniques are adopted, or else the country may find itself with a sophisticated tool that it cannot use.

(c) Available staff

The level of personnel resources that can be made available determines to a large degree how ambitious the data processing plan can be. The work of programming, maintaining, and running the system should not demand that programmers and analysts must constantly work overtime or feel continued pressure. It may be necessary to hire additional staff to meet the needs of the processing task, or to adjust the task accordingly.

The level of sophistication of the staff must be examined in addition to the number of people available. Processing data from a household survey demands a certain sophistication even in the simplest case. Training in programming languages, systems, design, editing concepts, and related topics may be necessary to augment previous training.

The level of sophistication of the data processing staff has a bearing on its ability to write customized software for editing, tabulation, and analysis. This is by no means the only factor that should govern the need for purpose-built software packages, but it is an important consideration.

Ironically, the majority of national statistical offices write their own customized software (or own generalized packages) for editing and to a lesser degree for tabulation, calculation of variances, and analysis. However, in light of the discussion presented in the next section, and with the emergence of more packages appropriate for statistical work, this is perhaps an area where countries can be encouraged to acquire and adapt what already exist. If purpose-built software can be identified to adequately accomplish the task at hand within the given constraints, then personnel resources that would otherwise be utilized in the development of such software can be devoted to other processing tasks.

C. Development of Custom Software versus Acquisition of Purpose-Built Software

There are at least four alternatives for providing needed applications software: in-house development of custom software, in-house development of purpose-built or package software, acquisition of existing packages, or commissioning custom-designed systems from software houses. The question is, when is it cost-effective to go outside as opposed to having existing staff develop the needed software? Such key issues as when the system is needed and how much money is available to spend on software must be addressed in answering this question.

In considering in-house software development, the data processing manager must determine:

(a) Whether the organization has existing staff to do the development, considering both quantity and quality.

(b) If development were done in-house with existing staff, how ongoing or parallel activities would be affected.

(c) If existing staff were insufficient, whether it is likely that additional staff could be hired locally for this task.

(d) If additional staff were hired, what would happen to those persons after the system development work is complete.

(e) How the computer usage required will affect the throughput of other work.

(f) How much continual manpower and machine time will be required to maintain the system (Wiseman, 1977, p. 18).

Staff time required to develop custom-coded programming systems must be heavily weighted, as a typical data processing centre in a statistical organization which does most of its computing via ad hoc COBOL, FORTRAN, or PL/1 programs may expend 30 percent or more of total staff effort in this activity. When considering salary and overheads, these programming costs may well exceed actual computing charges by a factor of 10 or more. It has been claimed that this expense could potentially be reduced by as much as 90 percent with the use of high level statistical languages available with prepackaged programs (Wilkinson, 1977, p. 309).

Proprietary software products, turnkey application systems, and data services extend the development capacity of organizations and increase their productivity. Ready-made software, in the form of a product or service, spreads development and maintenance costs over a broad set of users and also creates a community of interest and commitment that facilitates the software's validity and continued enhancement (Frank, 1979).

An organization that does not possess its own staff of experienced programmers capable of the timely development of valid and reliable applications systems may suffer serious consequences when attempting complex software development. Software becomes most critical as projects approach their expected target date for completion. When software is late or non-operative, the organization may incur additional expenses such as delivered but unused hardware, communications, facilities, and personnel. Parallel operations may also be jeopardized. Fortunately, this problem is avoidable when off-the-shelf software is available.

Arguments are quite compelling as to the obvious cost and performance benefits of utilizing available software systems distributed as a product. The discouraging facts, however, as reported in a 1977 survey of more than 300 companies, indicate that only 20 percent of all application software utilized was acquired externally (ibid.).

Standard arguments against acquisition of external software include:

(a) In-house staff would be offended at having to use programs designed by an outsider.

(b) Off-the-shelf software would preclude staff from the challenge and training of building the software itself.

(c) The "not-invented-here" syndrome, contending that each organization has its own unique problems and the capability to solve them.

(d) Outside software requires new training.

(e) Outside purchase can be matched by in-house development.

Many data processing shops in an attempt to justify in-house software development over acquisition of packaged software will use incomplete financial arguments because they ignore personnel overhead and computer costs incurred by the development team, as well as overlooking the cost of the definition, design, and maintenance phases. Also, in pointing the finger at generalized software's inability to meet stiff and inflexible in-house requirements, the user fails to realize that many of the special needs are more self-imposed than real, and even when real, are not often sufficiently important to justify considerable delays in completion of the task as a whole. Moreover, many application software products allow the user to insert customized code into the package.

Users must learn to identify and analyze the direct as well as the indirect benefits of externally acquired software products. The most significant benefit is avoidance of the classical in-house cost components. Indirect benefits include earlier operation, lower risk, saving key manpower and availability of more complete documentation.

The ultimate consideration in a make-or-buy situation is potential savings to the user. Analysis could be based on information displayed in a table such as the following:

	In-house	Off-the-shelf
Estimated development cost	$250,000	$ 35,000
Development time	15 months	3 months
Expected annual savings	$100,000	$100,000
Payback starts	18 months	6 months
Investment recaptured	54 months	18 months

This particular example would strongly recommend acquisition of the off-the-shelf product (ibid.).

It is true that statistical offices in several developed countries have invested heavily in the development of general-purpose software systems for use on a wide variety of their statistical applications. These systems have been necessary to meet their specialized requirements for survey processing, as well as to integrate a standard data management philosophy across all applications. This is not to say that these organizations could not have been adequately served by already existing software packages but rather, in most cases, that the decision to develop reusable systems was based on a specialized need and the availability of high level programming staff to do the development.

Most developing country statistical offices do not have the luxury of having an abundance of high level programming staff to be able to contemplate development of specialized reusable software packages. These organizations are encouraged to use existing software systems available from other statistical offices or vendors, and to integrate them with smaller customized routines required for special needs, in order to avoid the need to program one-time-use customized systems. Writing customized software would seem like the more risky approach to take, as data processing staff turnover is usually quite high in developing countries, making the maintenance of customized software virtually unmanageable.

The above does not imply that there can be no problems in the choice and operation of appropriate software packages, or that such packages are available to meet all or most of the needs of a continuing household survey programme.

D. Existing Software and Considerations in Adaptation to Developing Countries

As noted in the previous section, developing countries with a shortage of trained personnel and a restricted budget will find that there are definite advantages in acquiring packaged software for certain areas of survey processing such as editing and tabulation.

With the number of packages now available and the increasing sophistication of analytical techniques using ever larger data sets on diverse hardware, users at all levels are faced with a major problem in deciding which package to use. The problem is particularly acute for anyone who has to choose software to provide a service to a community of users as is the case in national statistical offices.

The use of an inappropriate software package can be devasting to producing correct results in an efficient manner. Certain sample designs preclude the use of generalized packages for statistical analysis. Other packages are not efficient for a large volume of data. The constraints of individual packages must be carefully scrutinized in the selection process and the temptation to use existing software even though it might be inappropriate must be avoided. The choice of the correct package for the task is essential if one is to take full advantage of packaged software.

This chapter discusses some of the considerations the potential user must weigh to assess adequately the appropriateness of any one package over another and to outline critical phases of software use for which the supplier should be able to provide support services to the user. Essentially the user will need to ask the following four questions:

(a) Capabilities: Was the package designed to help solve problems like mine?

(b) Portability: Can the package be transported conveniently to my computer?

(c) Ease of learning and using: Is the program sufficiently easy to learn and use that it will actually be useful in solving problems?

(d) Reliability: Is the program maintained by some reliable organization, and has it been extensively tested for accuracy?

A brief overview of the status of general statistical packages is presented in Annex I focusing mainly on highlighting those systems which have experienced the widest distribution to date in national statistical offices and pointing out available studies and reference books which describe, analyze, and compare many packages in greater detail.

1. Assessment of appropriateness

A major share of the burden of helping to make good decisions about the appropriateness of any one piece of software should rest

on the supplier of the software. The supplier, as the expert on the technical capabilities and limitations of his package, must work with the potential user to match the applications' needs with the features of the software. The supplier, of course, is going to emphasize the attributes of his system, especilly if the product is sold for profit. But what is important is that the supplier have experience in the field of intended application in order to be able intelligently to advise the user. For example, the vendor of business-use software most likely does not have an appropriate background to build software for scientific applications.

Beyond the initial inspection of a software package to determine if its capabilities meet the requirements of the specific application, the potential user must ask other key questions of the software suppliers and be prepared to do some investigating to verify the suppliers claims:

(a) Software design and implementation

-Is the design up to modern standards with a perceivable and coherent structure?

-Is the coding clear and explicit?

-Are conventions and standards adhered to?

-Can the coding be easily modified by the user?

-Is the software self-monitoring for ease of error diagnosis?

-Does a comprehensive benchmark test program exist?

(b) Transportability

-Is the system in a widely available language?

-Are there examples of specific installation projects?

(c) Support

-Has the system been successfully installed on a machine comparable to the user's machine and is its implementation guaranteed?

-Is there an established mechanism for notifying users of updates?

-Is there a users group?

-Is there documentation for implementers, maintainers, elementary users, advanced users, and user support staff?

-Are there teaching aids, such as audio-visual materials and sample data sets?

-Are courses available?

(d) Ease of use

-How often does the system fail on correct input?

-How easy is it to recover from incorrect input?

-Who are the target clientele?

-Is the system integrated to complementary systems? For example, are the output files generated by an editing system compatible as input to a tabulation system?

-How much does it cost to run?

(e) Statistical and numerical methods

-Are the methods robust and up-to-date?

-Are the methods suitable for the quality and quantity of data normally arising from a survey? (Rowe, 1980a, pp. 5-8)

Other key elements for assessment of appropriateness are briefly discussed below from the points of view of what a conscientious software supplier should be able to and is willing to provide.

2. Conversion

Before any statistical organization makes a commitment to acquire and use a software package for processing large or continuing surveys it must be determined if the supplier will be able to provide a version of the software for the existing hardware configuration. If a version of the software package is not currently available for existing equipment, a determination must be made about the time and cost involved in creating an appropriate version. Unfortunately, writing a software product in a "portable" language like COBOL does not necessarily mean that the amount of code modification necessary to make the package operational on different hardware is negligible. Even COBOL, which is considered the most portable language available, requires some careful scrutiny when trying to make sophisticated routines execute identically on different hardware.

In most cases, it would be more advantageous to commission the supplier to custom-convert the software to the target machine rather than having the user organization attempt an in-house conversion. No one knows the internals of a system better than its authors, and when paying for a converted product, the user receives a guarantee that the converted software will function correctly. If the supplier cannot readily put forth the resources to create an appropriate version, then the user organization must decide whether the necessary resources exist to do the conversion and whether the advantages of having that piece of software operational on in-house equipment outweigh the expense of doing the conversion in-house. The user must realize that in many cases suppliers provide more than adequate user documentation but fall short in developing and making available systems documentation that can aid a user in code modification.

One obvious indication of the portability of a package is the number of different versions that have been distributed to users since the initial release of the system. If a package has been on the market for some time and has not been converted for use on other machines, it is likely that the conversion of the package is a major undertaking and that the package contains a significant amount of machine-dependent or operating system-dependent code.

A commitment to statistical software packages should not force the user to lose flexibility in the choice of hardware which best meets the data processing needs of the organization. A supplier who has developed machine-independent software, and can support the continual conversion process for different kinds of hardware, provides that freedom of choice to the user.

3. Installation

One of the most complex, time-consuming, and critically important aspects of making a software system operational on any given computer is the installation phase. Not only must the software be compiled and stored in the resident libraries of the computer and benchmark-tested for accuracy, but it must also be fine-tuned to take full advantage of the computational and operational capabilities of the host configuration. In most cases, particular attention must be given to the location, size, format, and structure of input and output files to the system, as well as to the many intermediate files which are transparent to the user but which are vitally important as communication interfaces among program modules. Proper attention to the optimal peripheral device assignments, file recording techniques, and access methods can save substantial time and computer resources when the system is actually processing large volumes of data. The system must be integrated into the operating environment in such a manner as not to degrade performance of other vital processing which must run concurrently with this system.

The macro language procedures (e.g., sets of job control language) which are created to provide easy access to the appropriate part or parts of the package must be constructed in such a manner as to make the full operational capability of the system available to the user, but in a sequence which is logical and comprehensible to the user.

A software supplier can go a long way in helping the operations staff of the computer centre install the package in a way that will provide maximum benefit to the user community. Again, the supplier is most knowledgeable about the interaction of the program modules with various hardware devices and is in the best position to advise on the operational generation of the system. A supplier that has had previous experience with installations of his software on similar equipment should have available an installation guide or notes that highlight the most important aspects of successful installation on that hardware. If the package has been previously installed on similar equipment, the supplier should have maintained statistics on operational performance which could assist the new user in installing the package to operate as efficiently as possible in its new environment.

4. Maintenance

When assessing the software supplier's willingness and ability to provide the proper support to a user organization that commits itself to the acquisition and use of a software package, one cannot spend too much time investigating the supplier's perception of the magnitude of needed maintenance on the package and his past performance record in providing the necessary maintenance to other users.

It is recognized that maintenance costs can range as high as 70 percent of the total cost of system development and support. The supplier, therefore, should be in a position to expend a very large part of his total staff effort on the appropriate maintenance of the system he produces, especially if the software is being widely distributed and used by a diverse group of people. If it appears that the supplier is not in a position to offer adequate maintenance support but the statistical organization still desires to use that software because of its applicability, then that organization must be prepared to reserve substantial resources to provide adequate internal maintenance of the sytem to support the users throughout the organization. If this is the case, then there is much greater emphasis on the need for the software to be well tested and reliable and for the supplier to provide an abundance of clear documentation.

The user of statistical software should be aware of the diversity of activities in a full maintenance programme in order to more fully appreciate the importance of a supplier allocating

sufficient resources for maintenance activities. The types of maintenance activities which must be undertaken include corrective, adaptive, and perfective maintenance.

Corrective maintenance is performed in response to failures of the software. The most obvious type of failure is the processing failure, such as the abnormal termination of a program which forces job cancellation. These are attributed to "bugs" in the system. Failure of the software to meet performance criteria which have been specified in the system design is considered a performance failure. This case may not be a "bug". The problem may be caused by incomplete coding or failure to consider a feature of the hardware. Implementation failure is a third type of failure which may require corrective maintenance. A violation of programming standards or inconsistencies in the detailed design can lead to implementation failure.

Maintenance performed in response to changes in data and processing environments may be termed adaptive maintenance. Examples of change in the data environment would be a change in the classification code system associated with a particular element, or the logical restructuring of a data base. Examples of change in the processing environment would be the installation of a new generation of system hardware, necessitating recoding of existing assembler language programs; or the installation of a new version of the operating system, requiring modification of job control language statements.

Maintenance performed to eliminate processing inefficiencies, enhance performance, or improve maintainability may be termed perfective maintenance. Processing efficiency may be impaired by such things as an inferior computational algorithm, or inappropriate use of language features. It may be possible to improve cost-effectiveness of the performance by correcting these weaknesses. Performance enhancement may also be possible by making modifications, such as improving the readability of a report through reformatting, or adding a new data element to those included in a report generated periodically. Finally, although a program may be constructed and documented according to established standards, it may nonetheless be possible to improve its general maintainability. For example, a program may be made more readable through insertion of comments, or it may be made more accessible through rewriting its documentation (Swanson, 1976, p. 494).

When a statistical organization is "shopping" for software, particular attention should be paid to the cost of maintenance offered by the supplier. As has been outlined above, proper maintenance activities can often exceed the number of resources originally assigned to develop programs, so it should not be surprising to learn that the industry considers annual maintenance

costs to the user to run around 15 percent of the original purchase price of the software. Unfortunately, software companies have notoriously underpriced maintenance costs, usually charging 5 to 8 percent of product value. In the long run, this can only hurt the user (Frank, 1979).

A potential user should not rely solely on what the supplier promises to provide in the way of maintenance support. Too often the supplier's zealousness in "selling" his product to a new user far outstrips his ability to properly support the user. A better measure of the supplier's maintenance support is the comments from other users of the software. An experienced user is in a good position to relay his observations about the amount and quality of support offerred by the supplier and how much of his own staff time is required by the maintenance function.

5. Enhancement

The very nature of generalized software systems precludes their being able to handle a variety of applications in the most efficient manner. By choosing to use available software, the user has decided that the cost savings is worth the loss in flexibility and execution efficiency which will probably occur. However, inevitably some processing requirements are perceived by the subject-matter specialist as being so necessary and so inflexible that it may be necessary to modify the software package to meet the requirements.

For this purpose, many generalized systems provide easily accessible entry and exit points in the system whereby a user can insert a routine or set of routines which has been custom-coded and which will provide the precise calculation or speed factor required for critical procedures in the processing scheme. These user "windows" must be clearly documented by the supplier, spelling out the bounds within which the user may operate and the ramifications of misuse by the user. The supplier should have documentation that precisely states what kind of protection is built into the system to try to trap possible encroachments by user-inserted coding.

Not all software packages are designed to allow this user-coding interface. In some cases where it is allowed, the flexibility is not sufficient to meet the requirements of the survey application. In such cases, the user must define in what specific ways the software is deficient and determine the feasible alternatives for correcting this deficiency. Most software suppliers are keenly interested in continually trying to adapt their packages to reach an ever-widening base of users. In discussions with the supplier, the user may find that the supplier is more than willing to make slight adjustments to the software if they are

perceived as useful enhancements for other users as well. Other suppliers are very adamant about not introducing any permanent changes into a widely distributed system, but would be willing to work with the user to guide him in the best approach to modifying the package in-house.

6. Update documentation

Regardless of the age of any software package, it is inevitable that corrections, modifications, or enhancements will be made at some point in time. A supplier who properly supports his product will see to it that all users or subscribers will automatically receive notification of new releases of the software or updates to the basic documentation. In most cases, commercial suppliers charge an annual subscription fee for automatic updates in cases where the user has purchased the software. For users that lease rights to software, automatic updates are customarily included in the rental fee.

Non-commercial suppliers may not have an established distribution system for system updates, especially if their products were developed primarily for in-house use. If this is the case, the burden of assuring that the latest updates are in hand rests with the user. The user might ask the supplying organization to simply put him on a mailing list to be notified of any changes, and when updates are made the user can usually obtain them by reimbursing the supplier for the cost of reproducing materials and shipping them.

7. Exchange of information among users

The larger software houses either directly provide a service through which users of their software can exchange experiences and ideas or strongly encourage and support the organization of user groups. Most formalized programmes of user exchange evolve around a periodic publication distributed to member users which announces new releases of the software, problems encountered by individual users, and new innovative applications of the software by users. Some of the larger groups hold periodic meetings or conferences at which the vendor is invited to announce the latest developments and individual users are invited to present papers on successful applications of the package and newly developed routines to interface with the software (e.g., an assembler language routine written to read and format input data records more efficiently than the higher level language of the package).

If such organizations do not exist for a piece of software that is under consideration, the user should ask the supplier for a list of current installations using the product. In this way, an individual user can initiate his own information exchange with other users.

Unfortunately, to date there are very few established clearing-houses for the exchange of software products themselves for routines developed by individual user groups attached to a specific hardware vendor (e.g., UNIVAC USE Program Library Interchange at the University of Wisconsin). The reasons for the lack of such software exchanges are quite evident. Beyond the problems of defining what the exchange environment should be and what organizational body or bodies should administrate it, more profound legal considerations exist, such as protection of proprietary rights, protection of the exchange supplier against liability for misuse, and protection of the user against claims of unauthorized possession. Very basic agreements would have to be reached among all parties, whereby the suppliers would have to relinquish all claims to the software and guarantee that any software submitted by any user to the exchange contain no proprietary code.

8. User interface with supplier

It suffices to say that the potential user of a software package cannot make a judgment about whether or not to invest resources in the use of that package without thoroughly investigating the supplier's position on the points made above. In many cases a national statistical office initiating a continuing programme of household surveys will be seriously reviewing many software packages for the first time. If the statistical office is leaning toward the purchase of new software from a commercial vendor, then a few words of caution would be appropriate.

The user should not expect:

(a) The selection process to be a light chore.

(b) To receive as much attention from the vendor after the sale as before.

(c) The package to handle everything in the most efficient way.

(d) That having the package will necessarily guarantee meeting an initial implementation schedule.

(e) To receive anything from the vendor other than what the vendor explicitly agreed to provide (Gantt, 1979, p. S/4).

The prospective software user must be prepared to discuss frankly the advantages and disadvantages of a supplier's piece of software with the supplier representative. The following areas are critically important to cover with that representative:

(a) The supplier should be aware of the user's hardware configuration.

(b) The supplier should openly inform the prospective user of any hardware innovations needed.

(c) The user should ask for a product demonstration in a similar environment to that of the target machine.

(d) The user should review the supplier's pricing agreements, if any.

(e) All interested parties in the organization should attend the supplier's presentation.

(f) The user should discuss and compare competitive products.

(g) The supplier should leave sufficient technical information about the product behind in order to enable the user to make a sound decision (Datapro Research Corporation, 1978, p. 26).

9. Training requirements

A very important aspect to the successful integration of a specific software package into the overall processing system for a household survey programme is the availability of proper training programmes for the users as well as of training and reference materials. All too often a decision is made to use a certain package based on the merits of the system described in a manual or a brochure without properly investigating the availability of training programmes and materials from the supplier.

At the very least, a software supplier should have some mechanism available to provide formal training on the installation, use, and maintenance of software packages. Without assurances from the supplier that training can be made available, it would be unwise for a statistical office to acquire an unfamiliar package and expect to be able to use it successfully.

A training programme can be approached in one of two ways. The optimal situation would be having the software supplier send an expert to the installation to train the entire group of users at the site. In this way the package can be demonstrated in the environment in which it will actually be used and the trainer can customize the presentation to the specific applications planned.

If this type of on-site training is not possible, then an alternative is usually available. Most commercial software vendors

offer courses periodically at their headquarters or regional offices and the statistical office could send a technician to that training with the understanding that this person would be responsible for training other users back in the office. However, apart from being less convenient and confined at best to a few representatives from the user's organization, these courses in many cases tend to "be off-the-shelf" and not designed to teach the use of a package with respect to the specific applications and conditions of any one organization. Some commercial vendors will provide instructors to teach at a user's facility and tailor the course to the organizations's needs, but the costs involved are often prohibitive.

Apart from formal training courses, the availability of adequate documentation and training materials is also crucial. In some cases, software is delivered to the user with a set of documentation that includes a user's guide and other reference materials which will help the novice user in his initial attempts to use the system. A few software houses have even developed structured, self-teaching manuals geared toward a new user learning the fundamentals of the system at his own pace. However, the vast majority of software packages are supported only by technical reference materials which are not aimed at instructing a user having no previous knowledge of the package.

Software developers from non-profit organizations, such as government agencies and universities, may tend to have more training materials available which have been geared to teach the use of their systems for the particular applications of the organizations to which they belong. These materials are usually available for distribution with the software when it is delivered. However, these organizations are less likely to conduct periodic courses for outsiders, as the primary purpose of their software is in-house use.

Organizations that develop generalized software primarily for export overseas are usually better equipped with training materials and available staff to provide individualized training to outside users. Institutions such as the United Nations Statistical Office, the International Statistical Programs Center of the United States Bureau of the Census, the World Fertility Survey, the Overseas Development Administration in the United Kingdom, and a few select commercial vendors that heavily concentrate on overseas markets have existing programmes for delivering and teaching their software products. In addition, some donor government agencies offer regional training at a host site where representatives of many countries in that region can attend software training. An example of this is the two year programme initiated by the United States Agency for International Development to provide a private contractor to develop and teach a series of regional workshops on the use of the edit package COBOL CONCOR for processing housing and population census data.

Any statistical office must be prepared to set aside ample time for programmers to be fully trained in the use of a new package. Just as one cannot expect a programmer to become fully knowledgeable in a new programming language in a couple of days, neither can one expect a person to be trained in the use of a complex package in a few days. Training may very well range from a few days for simple systems to many weeks for complex systems. This commitment to training must be viewed as a wise and necessary use of resources.

E. Technical Assistance and Training

Many countries will require technical assistance in establishing a capability for data processing for continuing household survey programmes. In certain more advanced countries, a short-term consultancy to provide assistance in specific specialized areas such as installing new hardware or software or setting up a new system may suffice. But, in a majority of the countries participating in the NHSCP, especially countries undertaking regular household survey activity for the first time, long-term technical advisory services will probably be necessary. One of the important functions of any technical adviser must be to assist the organization in organizing formal and informal training for its staff, including on-the-job training of counterpart staff.

The importance of training has been stressed earlier in this document (see sections III.B.2 and VI.D.9). Training requirements, procedures and facilities, including those in the field of data processing, will be discussed in a forthcoming technical study of the National Household Survey Capability Programme.

Information on training courses in data processing offerred by various international and regional training institutions is presented below in Annex II of the document. Requests for additional information or queries may be directed to the United Nations Statistical Office or statistical divisions of the regional commissions.

VII. CONCLUDING REMARKS

The processing of data from continuing programmes of household surveys requires considerable facilities and skills. As was stated at the beginning of this document, there can be no "packaged" approach to this undertaking. Instead each country must develop its own course of action in accordance with its needs and means of meeting them. This study has discussed the various factors involved in making appropriate choices, rather than recommend any single approach. At the same time, its objective has been to promote good practices in the design and implementation of procedures for statistical data processing.

In conclusion, some elements essential for the success of the data processing effort are:

(a) Insistence that the data processing be kept to a manageable task.

(b) Good communication among the data processing staff, the sampling specialists, and the subject-matter specialists.

(c) Realistic planning of all facets of the data processing effort.

(d) Accurate assessment of the existing data processing staff, hardware, and software and an effort to augment or improve them as necessary.

(e) Sound system design and complete testing of all software.

(f) Careful control during production processing.

(g) Complete documentation.

And some practical advice to data processing managers and experts:

(a) Select one person or a committee from the data processing staff to participate in setting goals and planning.

(b) Keep budget, resources, and schedule in mind as planning proceeds and do not hesitate to indicate potential data processing problems.

(c) Participate in questionnaire design to assure processability.

(d) Do a comprehensive system design, taking into account volume and timing.

(e) Decide whether or not current staff can handle programming, production processing, and maintenance responsibilities. If not, hire additional staff and train current staff.

(f) Decide whether or not existing data entry equipment and computer hardware are adequate for the processing task. If existing equipment is inadequate or there is no access to equipment, begin the procurement process as early as possible.

(g) Decide whether or not additional packaged software is needed. If so, carefully study the packages available before making a choice.

(h) Make arrangements for providing training to existing and new staff, especially in the use of new software packages.

(i) Work with sampling and subject-matter specialists to develop detailed specifications for all computer programs to be written. All specifications should be in writing.

(j) Be sure all software is thoroughly tested prior to production processing. This includes review of output by sampling and subject-matter specialists to be sure the programs meet their needs.

(k) Apply effective systems of quality and operational control during production processing.

(l) If problems arise, do no try to hide them, but rather attempt to deal with them in a straightforward manner that minimizes their effect on the budget and the schedule.

(m) Maintain complete documentation of the system of programmes and of production processing.

(n) Learn from mistakes made in processing one round of the survey, so that the next round may be improved.

(o) Seek technical assistance in any area where the need is indicated.

ANNEX I

A REVIEW OF SOFTWARE PACKAGES FOR SURVEY DATA PROCESSING

A. Introduction

1. Criteria for evaluation of available packages

National statistical agencies in developing countries will in most circumstances find it necessary as well as advantageous to utilize existing software packages when possible, rather than to devote their scarce personnel and budgetary resources to developing new software. This is true at least of certain areas of survey processing such as editing and tabulation.

However, with the large number of packages now available and the increasing diversity of hardware, size of data sets to be processed and sophistication of analytical techniques, users at all levels are faced with major problems in deciding which packages to choose. The problem can be particularly acute for an agency responsible for providing data processing services to a community of users, as often in the case with national statistical offices.

This Annex provides a review of the requirements, capabilities and limitations of the major packages which national statistical agencies might find useful in survey data processing. The objective of this discussion is not to attempt an exhaustive enumeration of all the available software packages to recommend a select few; rather, the objective is to identify the discrete tasks involved in computer processing of household sample survey data and to try to identify some useful software packages which are appropriate for this type of processing. Certain specific packages are included in this review because of their frequent use by or easy accessibility to national statistical agencies; others mentioned may be less widely known or used but offer promise in the various areas of statistical processing.

There are a number of good sources (listed at the end of this section) which provide comprehensive inventories of available software products and rate them according to standardized criteria. However, most publications dealing with the description, classification, and evaluation of statistical software packages view such software as being used in an academic environment or by a subject-matter specialist. The criteria used to evaluate these products lean heavily toward measuring how well a statistician can learn and use a given product individually with little or no assistance from the data processing staff. These may indeed be the most relevant criteria in certain environments - when, for example, the statistical analyst has no access to data processing

professionals but does have access to a computer; or when the subject-matter specialist with data to process has difficulties in communicating with programmers and perceives that dependence on the latter will result in serious bottle-necks and delays. This however, typically is not the environment in a large statistical organization engaged in regular and relatively voluminous collection and processing of data. The real objective for a national statistical agency in choosing a piece of software has to be to minimize the resources required to process a given set of data, utilizing all available facilities including the services of professional data processing staff. Apart from suitability for the task, other requirements in the choice of particular software are its portability and availability, and the documentation, maintenance and training support provided by the supplier. It is on the basis of these criteria that the following review is undertaken.

The selected packages have been grouped according to their major function in survey data processing. The major functions considered are:

(a) Editing, such as interactive data entry, structure, range and consistency checking, error reporting and manual or automatic correction.

(b) Tabulation, including the computation of means, medians, percentages, which these tabulations may require; printing of tables, particularly in a photo-ready form.

(c) Computation of sampling variances and co-variances.

(d) Survey analysis, such as fitting linear and log-linear models, multivariate and cluster analyses, various types of statistical tests, and general data and file manipulation.

(e) General statistical programmes, which are distinguished from group (d) only for having much wider capabilities for statistical analysis.

(f) Data management, such as matching of files, extraction of records, manipulation of data arrays and data retrieval.

Generally, particular statistical packages have more than one function; the classification adopted here is according to what is considered as the primary function. For example, RGSP is listed here as a "tabulation" package, while it is labelled by its developers as a "general package for survey analysis". Most packages have some editing or data validation facilities, as well as capabilities to recode or reformulate data. In fact, recoding of raw data, which is an important step in household survey data

processing, has not been identified above as a separate function. Packages used for certain other functions, such as survey design and sample selection, have been developed by particular statistical organizations but are not included here for being too specific in function and use.

Within each functional group packages may be distinguished according to the degree to which they are portable and the extent of their use in national statistical organization. Portability depends upon the language(s) in which the package is written, the degree of its machine independence, core storage and interface requirements, and the variety or environments in which it has been successfully installed. How widely a package is used depends upon its portability as well as the quality of support (documentation, installation, training and other assistance) provided by the supplier.

The level and design of the user language interface to the package is a good indicator of the overall quality and usefulness of a package. Users have a right to expect to be able to describe their data and the operations to be performed on that data in a clear, concise, and intelligible form and in statements free from extraneous technical details. A well-designed language simplifies the task of providing user documentation and is the key to properly modularized computer implementation. The more general and flexible the language, the greater the frequency of use of a package and hence the greater the incentives for the developer/supplier for refining and extending the system and achieving portability (Wilkinson, 1977, pp. 229-300). There appears to be a consensus of software evaluators that, at the highest levels, the problems of generalized computer systems design are exclusively those of language design. It is perhaps no accident that packages that evaluators find to be both powerful and simple to use were designed and implemented by people who stress the language approach (Francis and Sedransk, 1976, p. 2).

2. List of packages reviewed

A brief review of the requirements, capabilities and limitations of the packages listed below is provided in the subsequent sections. The rating of each package according to the criteria of portability and use described earlier is also identified as follows:

- **** Widely used packages with high degree of portability
- *** Widely used packages, but with limited portability
- ** Packages used less widely in statistical offices, but which show promise
- * Other packages with restricted distribution and portability

Editing programmes	Portability
COBOL CONCOR	****
UNEDIT	***
CAN-EDIT	*
Tabulation programmes	
CENTS-AID III	****
COCENTS	****
RGSP	****
LEDA	***
TPL	***
XTALLY	***
GTS	*
TAB68	*
Survey variance estimation	
CLUSTERS	***
STDERR	**
SUPER CARP	*
Survey analysis programmes	
GENSTAT	****
P-STAT	****
FILAN	**
BIBLOS	*
PACKAGE X	*
STATISTICAL ANALYSIS	*
General statistical programmes	
BMDP	****
SPSS	****
OMNITAB-80	***
SAS	***
Data management programmes	
CENSPAC	***
EASYTRIEVE	***
SIR	***
FIND-2	*
RAPID	*

3. Source of further information

The major sources of information pertaining to the existence, scope and availability of statistical software include:

(a) A Comparative Review of Statistical Software, Exhibition of Statistical Programme Packages, New Delhi, 1977, edited by Ivor Francis for I.A.S.C.

(b) Statistical Software: A Comparative Review for Developers and Users, by Ivor Francis and Lawrence Wood, El Sevier North Holland, New York, 1980.

(c) Statistical Software for Survey Research, prepared by Beverley Rowe for Study Group on Computers in Survey Analysis, World Fertility Survey, London, 1980.

(d) Statistical Computing Enviroments: A Survey, Australian Bureau of Statistics, February, 1979.

Several professional societies have formed committees to evaluate software. These include:

(a) American Statistical Association (A.S.A.).

(b) International Statistical Institute (I.S.I.).

(c) Similar groups found in New Zealand, Sweden, and Japan.

Many conferences are held each year or biannually which invite technical papers on the evaluation of statistical software. They include:

(a) International Biometric Conference.

(b) American Statistical Association.

(c) Symposium on the Interface of Computer Science and Statistics.

(d) International Statistical Institute.

(e) New Zealand Statistical Association.

(f) Symposium on Computation Statistics.

(g) International Association of Mathematical Geology.

(h) INTERFACE, an annual North American conference.

(i) COMPSTAT, a biannual conference now organized by the International Association for Statistical Computing (IASC).

B. Editing Programs

1. COBOL CONCOR version 2.1 (Consistency and Correction system)

This package was developed and is distributed by the International Statistical Programs Center (ISPC) of the United States Bureau of the Census, Washington, D.C. 20233, U.S.A. ISPC fully supports the package abroad, providing workshops on its uses, as well as technical consultation and trouble shooting. A comprehensive set of documentation is available, including a technical reference manual, a systems manual, and a diagnostic message manual. A user's guide developed by NTS Research Corporation of Durham, North Carolina, U.S.A., is also available.

The package is a special purpose statistical software package which is used to: identify data items that are invalid or inconsistent; automatically correct data items by hot-deck or cold-deck imputation; create an edited data file in original or reformatted form; create an auxiliary data file; produce an edit diary summarizing errors detected and corrections made; and perform error tolerance analysis.

CONCOR can be used to inspect the structure of a household questionnaire, the validity of individual data items, and the consistency among items both within a logical record and across logical records within the same questionnaire. The system generates error messages as the data are being inspected based on the validity and consistency rules set forth in the user's programme. The user can supplement CONCOR's message system by supplying any specifically desired messages which will be displayed as errors are detected. Error messages can either be generated on a case-by-case basis or summarized over any desired area. The program displays the frequency with which data items have been tested, the frequency of errors found, and the error rate. These statistics can be displayed for the total run or for specific disaggregate levels. Tolerance limits can be set by the user and if they are exceeded, the system will reject a defined work unit as being unacceptable.

Corrections to data can be imputed using hot-deck arrays, cold-deck arrays, or through simple arbitrary allocations. The system maintains counts of the frequency with which the original values of data items are changed.

The system produces an edited output file identical in format to the unedited input file, which allows the output to be treated as input and read back through the system to determine if changes made during the editing process have introduced any new inconsistencies. A derivative output file can be produced concurrent with the edit run. The command language is free-format in design and the system provides a comprehensive syntax analysis function which protects the user against coding errors and execution-time errors.

The system can read files produced by many other packages and its outputs are compatible to the COCENTS and CENTS III tabulation systems. However, the system is presently restricted to handling only fixed-length records. The package is written in low-level ANSI COBOL and is comprised of 19 COBOL source modules. It requires 128K bytes of primary storage, as well as 4 million bytes of on-line storage. Versions of the system are installed on IBM OS and DOS systems, HONEYWELL 66, ICL 2970, UNIVAC 1100, NEC 500, WANG VS80 and Perkin-Elmer 3220.

2. UNEDIT

This package has been developed and distributed by the United Nations Statistical Office (UNSO), New York, N.Y. 10017, U.S.A. The system is a generalized edit package developed to meet the needs of census and survey editing on small computers. The package requires only 32K bytes of primary storage and 5 million bytes of fixed-disk storage.

Rather than a command language, the user codes a series of parameter-like statements which can perform the following functions:

(a) Identify data invalidities (out of range values).

(b) Identify intrarecord and interrecord inconsistencies.

(c) Perform arithmetic calculation and comparison of data fields.

(d) Perform analysis of edit rules to ensure consistency and identify implications.

(e) Check structure for missing records.

One of the advantages of UNEDIT is that it can process hierarchical files, as well as flat files with multiple record types. Error statistics by type of error and by data field name are printed. No capability exists for automatic imputation, though arbitrary assignment of a value to a data field can be done under

certain circumstances. The system does not have the capability to display statistics based on weighted estimates (which would be of value when editing sample data), nor does it have a tolerance check to indicate the number of changes introduced into the data.

The system consists of two modules, one for pre-edit preparation, the other for the execution of editing. Both modules are written in RPG-II, and therefore the number of machines on which they could be installed is somewhat limited.

The UNSO fully supports UNEDIT. The package is easy to install; in fact UNSO has had success in installing the package in statistical offices overseas by simply putting it in the mail. A COBOL version of UNEDIT is now in the test stage.

3. CAN-EDIT (alias GEISHA - Generalized Edit and Imputation system using Hot-deck approach

This package was developed by Statistics Canada, Ottawa, Canada. This system for automatic edit and imputation has been implemented in a data base environment and is used in household surveys and population census processing. A high-level, non-procedural language set is used to specify the editing rules, which are expressed in the form of a set of conflict statements. These conflict statements can be directly specified by subject-matter specialists and can be fed into a specification subsystem which analyzes the edit rules and lists possible contradictions, redundancies, and implications that are inherent in them. The system provides a summary of the number of records which fail different conflict rules or combinations of rules.

Imputation requirements are determined directly from the edit rules and are based on two criteria:

(a) The specified edits should be satisfied by making the smallest number of changes in the data.

(b) Frequency distributions of values in the data should be maintained to the greatest degree possible.

The system retains both the imputed and unimputed data to assess gross and net changes introduced.

The package presents one of the most powerful and comprehensive edit processors available but its usefulness to developing country offices is very limited. It is written in PL/1

and Assembler and requires IBM 370 equipment having 200K bytes of primary storage and the entire survey file loaded on direct access devices. The system is tied into the RAPID data base management system developed by Statistics Canada, which means that RAPID must also be installed on the target IBM computer if CAN-EDIT is to be implemented.

C. Tabulation Programs

1. CENTS-AID III

This package was developed and is distributed by Data Use and Access Laboratories (DUALabs) Arlington, Virginia, U.S.A. The developers of the package describe it as a high speed computer system engineered to minimize the cost of processing large data files through the use of generative programming technology. (Generative means that the system interprets a set of user-supplied parameters and builds an executable programme tailored to the user's request). The system allows the user to generate and display cross-tabulations of up to eight dimensions and produce percentages, means, medians, standard deviations, variances, and chi-squares for any table produced. The system can also transform and recode data and provides report formatting commands.

Data records of fixed or variable length can be processed, as well as any file containing up to 26 different record formats. Hierarchical data structures of up to 30 levels are supported. The system obtains its technical information and descriptive labels for data variables from a computer-readable code book called a Data Base Dictionary.

Beyond producing cross-tabulated reports and related survey statistics, the package can produce subfile extracts; generate and display correlation and co-variance matrices; and create an SPSS Correlation Interface File.

The system consists of seven programmed modules in ANSI COBOL and it makes use of a utility sort. It requires 168K bytes of primary storage, as well as 40 cylinders (300K characters) of 2314 disk space for temporary work files and three cylinders (22K characters) of permanent on-line storage.

DUALabs provides software support to users of the CENTS-AID system in the form of training classes, consultation for user problems, and software updates for system errors (Hill, 1977, P. 229).

2. COCENTS (Cobol Census Tabulation System)

This package was developed and is distributed by the International Statistical Programs Center (ISPC) of the United States Bureau of the Census, Washington D.C. 20233, U.S.A. ISPC fully supports the use of COCENTS as well as its companion package CENTS III overseas, providing workshops on its use and technical consultation and trouble shooting of system problems. It is probably the most widely used tabulation software in national statistical offices.

The system is comprised of five program modules written in ANSI COBOL and requires 64K bytes of primary storage. It can read files produced by many other packages, including the edit package COBOL CONCOR.

COCENTS is a special purpose statistical software package which is used to manipulate data files, cross-tabulate individual observations, aggregate tabulations to higher levels, perform simple statistical measures, and format a publication quality tabular report. It can be used to extract or select certain subuniverses or samples from a data file or to recode individual data items prior to tabulation.

The approach to the tabulation of data involves the preparation of tally-blocks for the smallest observational units or areas desired, the consolidation of these tally-blocks in report form. The system can operate on complex hierarchical files, which are common to household surveys, and it can also operate on flat files. However, the system requires that all records be of fixed length. Observational units can tabulated, weighted or unweighted, and new variables can be defined by grouping or reordering. Publishable tables can be produced from a data file in one run, allowing the user full control over the appearance of these tables. Basic statistics can be produced, including totals, subtotals, percentage distributions (to one decimal place), ratios, means, and medians.

The user instruction set provides a great deal of flexibility, but in doing so is more oriented toward use by programmers than by subject-matter personnel. For statistical offices engaged in continuing programmes of household surveys, a major advantage of the package is its capacity to produce well laid out tables ready for immediate publication. On the other hand, individual tables require elaborate coding, and any modification to existing tabulation programmes can be tedious. To overcome this difficulty, the World Fertility Survey, International Statistical Institute, London, developed the programme COCGEN which acts as a preprocessor to and generates parameter cards for COCENTS.

COCENTS is undergoing extensive redevelopment and both it and its companion CENTS III will be replaced by a single package, CENTS IV. The new package is expected to provide significant improvements in the areas of self-documenting, structured system source code, totally free-format user language, enhanced error protection and error message system, more flexible display capabilities, and more power in the language.

3. RGSP (Rothamsted General Survey Program)

This package has been developed and distributed by RGSP Secretariat, Computer and Statistics Departments, Rothamsted Experimental Station, Harpenden, England. It is labelled by its developers as a general package for survey analysis. More specifically, it is used for the formation, manipulation, and printing of tables from survey data. The system is divided into two parts: Part 1 is a FORTRAN subroutine package which forms the tables and Part 2 performs table manipulation and printing.

The system can manipulate survey data to some extent before they are entered into a table, but this is limited to: allowing for missing values, blank fields, and invalid punches; conversion of alphanumeric codes to numeric values; and exclusion of erroneous values from tables. If true edit detection, error reporting, and correction procedures are to be employed before creating tabulation, then these algorithms must be included in a custom-written FORTRAN programme which in turn calls the FORTRAN subroutines that produce the tables. The more powerful part of the system, Part 2, provides for the following table manipulations:

(a) Addition, subtraction, multiplication, division.

(b) Extraction of square roots.

(c) Percentage distributions.

(d) Combination, reordering, and omission of table levels.

(e) Creation of subtables from tables.

(f) Combination of tables or parts of tables into new tables.

(g) Calculation of ratio estimates and regression estimates.

(h) Calculation of standard errors.

A very important feature of this system is its ability to handle hierarchically structured data; it can also produce standard errors for stratified, multistage, clustered samples. The capability to handle this kind of data structure and sample design,

which are common to household surveys, makes RGSP a versatile package. For analysis such as multiple regression and fitting of models to multifactorial tables, the package provides interfaces to other Rothamsted packages, such as GENSTAT and GLIM. The system requires 128K bytes of primary storage and a FORTRAN compiler. Versions exist and are supported for ICL, IBM, NCR, and CDC machines. For further description see the RGSP users' guide and Yates (1975).

4. LEDA

This package was developed and is distributed by the Institute National de la Statistique et des Etudes Economiques (INSEE), Paris, France. The developers of the package call it a survey analysis system. It is organized into three major operations, each of which is handled by a separate program module.

File building and manipulation are achieved through the CASTER module. The building of a file is comprised of identifying hierarchical relationships of records in a tree structure. The file is compressed so that data variables common to lower hierarchical levels are represented only once at the upper level. A dictionary naming all variables is defined. Range checks can be made on individual items, as well as file structure checks on the interview. Automatic rectification via hot-decks can be applied.

Editing is accomplished through the POLLUX module. This program performs logical or consistency checks between items and also provides file maintenance functions, such as record deletion, recoding transformation of structures, and subfile processing.

Tabulation is the responsibility of the POLLUX module. This program provides for the tallying of variables defined in the data dictionary or newly created variables. It allows for definition of restricted universes or subpopulations on tables and displays totals and percentages. An important feature of this tabulation phase is its ability to handle fractional numbers in floating point arithmetic. It also allows for user-defined computations to be performed on tables produced.

Although the LEDA system generates a COBOL program for execution, it's software is written in CPL/1 and Assembler. This limits its portability, as it is currently only operational on IBM 360/320, Honeywell Bull IRIS 80, and Honeywell Bull 66. The system requires 160-192K bytes of primary storage. The command language can be written in either English or French, and documentation is available in both languages. The system is supported by INSEE and is under active development.

5. TPL (Table Producing Language)

This package was developed and is distributed by the Division of General Systems of the United States Bureau of Labor Statistics (BLS), Washington D.C. 20212, U.S.A. (Distribution in Europe is handled by the International Computing Center in Geneva, Switzerland). The system was created as a computer language to produce statistical tables. It can cross-tabulate, summarize, and use the results for statistical and other arithmetic calculations. New variables can be defined by grouping, deleting, or reordering. Publishable tables can be produced in one run, with great format flexibility available to the user.

A CODEBOOK is used to describe the attributes of the data variables to be tabulated. Once the CODEBOOK is established, the user can code table statements in a very English-like manner to produce the desired reports.

The package can handle complex hierarchical files, as well as extremely large numbers (10 to the 75th power) with 16 significant decimal places. Fixed as well as variable record formats are allowed, and input data can be in floating point, binary, character, or packed decimal forms.

TPL receives high marks from most software evaluators for the simplicity and power of its command language. It also allows the user flexibility of having the system automatically format a table with little user coding required. It is not necessary for the user to submit precisely coded instructions to control every aspect of the table's appearance.

The package can produce basic statistics, including percentages, means, medians, quantiles, and standard deviations.

The major drawback of the system is its size and implementation language. TPL is written in XPL and machine language, which restricts its use for the most part to IBM computers. It also requires 300K bytes of primary storage and approximately 6400 tracks (84 million bytes) of 3330 disk space. Nevertheless, TPL is one of the most widely used tabulation systems.

TPL has a user's guide containing many examples, and all diagnostics generated by the system are English language messages. BLS offers courses in TPL in Washington, but the developers claim that many persons have used the package successfully without taking the course.

6. XTALLY

This package was developed outside the United Nations, but has been distributed by the United Nations Statistical Office (UNSO), New York, since 1976. The system has been used to tabulate census, survey, and administrative data in many UNSO projects in developing countries. It is capable of producing multi-dimensional cross-tabulations summing one or two variables or counting records and giving subtotals, percentages, ratios, means, differences, sums, or weighted totals at all levels of tabulation.

XTALLY tabulates through disk stored fixed arrays, whose segments are selectively brought in and out of primary storage using a primary stored buffer for partial accumulations that enables minimal swapping of array segments. The trade off for this capacity is speed, as XTALLY operates substantially slower than some other tabulation systems, which is not significant if the system is being used to tabulate small household sample surveys.

The package can handle tables of up to seven dimensions, each containing up to 126 classification categories. The system uses a data dictionary to predefine data record variables. The user parameter language is simple and straightforward. Only three statement formats are used and they can be learned by the non-programmer in a few hours. The system has little flexibility in the specification of tabular report format, so it should not be construed to be a package that can readily create publication quality output. Since the system is an interpretive system (i.e., it reads the user's parameters and executes appropriate modules of an existing programme), it does not require any compilations by the user.

XTALLY lacks total portability, as it is implemented in the RPG-II language. It does, however, fit most small machines, requiring as little as 24K bytes of primary storage and two million bytes of disk storage (Lackner and Shigematsu, 1977, pp. 273-274). A COBOL version of XTALLY, produced by a United Nations data processing expert on field assignment in Africa, is ready for testing.

7. GTS (Generalized Tabulation System)

This package was developed and is distributed by the Systems Development Division of the United States Bureau of the Census, Washington D.C. 20233, U.S.A. It consists of a series of computer programs designed to produce statistical tables economically. The system is controlled by an English-like command language and previous experience with computers and programming languages is not a prerequisite. A user should have a basic understanding of the terminology and concepts used to describe tables and data.

The command language is very structured and powerful. It references a user-defined dictionary. GTS was designed to be one module of an integrated generalized statistical system and was developed to meet the following five objectives:

(a) Bridge the conflict between being easy-to-use and powerful.

(b) Function in a conversational as well as a batch mode.

(c) Exploit the availability of large core storage on the UNIVAC 1100.

(d) Maintain consistency in recoding of the input data.

(e) Maintain flexibility without loss of machine efficiency.

The system provides great flexibility in the user's ability to construct, manipulate, and format tables of a publication quality. It provides for the creation of means, medians, ratios, percentages, and square roots, and has capabilities for handling various kinds of survey weighting schemes and data formats.

Even though GTS is written in ANSI COBOL, the system is very dependent on UNIVAC 1100 computers, as its I/O drivers use custom FORTRAN subroutines to handle special data formats. The system is quite large, taking full advantage of the tremendous amount of storage available on the UNIVAC 1100's. The Bureau of the Census fully supports all internal users of GTS but does not plan to convert the system for use on other machines or support its use by other organizations.

8. TAB68

This package was developed by the National Central Bureau of Statistics in Stockholm, Sweden. It is a programming language for table creation, which is structured into simple primary and secondary key words. Only input data and output data need to be described. The package can create frequency tables, summation tables, and percent tables. Available information does not give specifics of table design or flexibility. Existing documentation in English consists of a handbook and a reference card. The system is written in IBM Assembler language and is installed on IBM 360 OS and IBM 370 MVS computers. No information was available on other computer requirements.

A companion package to TAB68 is a package for record linkage called STRIKE. Up to 30 sequential input files may be matched in one run. An unlimited number of output files may be produced by

coding a few English primary and secondary key words. The STRIKE system is available for IBM 360 OS and 370 MVS computers.

D. Survey Variance Estimation Programs

1. CLUSTERS

This package was developed and is distributed free of charge by the World Fertility Survey, 35-37 Grosvenor Gardens, London SW1 OBS, United Kingdom. It has been distributed to 30 institutions participating in the WFS programme and to 10 additional sites. The program is written in FORTRAN and requires approximately 50K bytes of core memory. It has been installed on IBM, ICL, CDC and Hewlett-Packard equipment.

CLUSTERS computers sampling errors taking into account the clustering, stratification and other features of the sample design. The sample data may be weighted, and the program will handle many different sample designs. The statistical approach for computing standard errors is a first order Taylor approximation.

The program reads data according to a user supplied FORTRAN format statement. The data must be in the form of a "rectangular" sequential file with no non-numeric characters in the field being referenced. Comprehensive recode instructions are available for creating derived variables and for selecting subclasses or subpopulations of the data. Sampling errors for descriptive statistics are computed, including proportions, means, percentages, ratios, and differences in these. In addition to standard errors, CLUSTERS produces two derived statistics: The Design Effect and the Rate of Homogeneity. They provide the basis for generalizing the computed results to other variables and subclasses of the sample.

The sample structure can be specified in a flexible manner, either as data fields coded on individual records or as separate parameter cards. The computations to be performed are specified as a matrix of substantive variables (which may be recoded from existing data fields) and sample subclasses or subgroups. For each variable, sampling errors are computed for the total sample, for each of the specified sample class and for differences between pairs of classes. Furthermore, the total sample can be simply divided into a number of geographical domains, and computations for all variables and subclasses repeated for each domain. This makes the program suitable for large scale and routine computation of sampling errors which may be of considerable value in the development of survey designs for continuing programmes of household surveys. For further information, see User's Manual (Verma and Pearce, 1978).

2. STDERR

This package was developed and is distributed by the Research Triangle Institute, P.O. Box 12194, Research Triangle Park, North Carolina 27709, U.S.A. STDERR computes certain ratio estimates or totals and their standard errors from the data from a complex multistage sample survey. The sampling units at various stages may be drawn with equal or unequal probability and may be stratified.

The ratio estimates and their standard errors are computed for various domains of the population. Standard errors for the estimated differences between the domain estimates and the estimates for the entire sample population are also computed. The statistical approach used for computing the standard errors is a first order Taylor approximation of the deviations of estimates from their expected values. This program gives one of the best known feasible approximations in currently available literature of standard errors for a large number of ratio estimates.

The program is written as a SAS procedure. The syntax of statements is similar to SAS and the user may make any data transforms using SAS data statements.

STDERR is available for IBM and IBM-compatible machines. It is currently installed at 25 sites.

3. SUPER CARP

This package was developed and is distributed by the Department of Statistics, Iowa State University, Ames, Iowa 50010, U.S.A. SUPER CARP is a package for the analysis of survey data and data subject to measurement errors. Is is capable of computing the co-variance matrices for totals, ratios, regression coefficients, and subpopulation means, totals, and proportions. For subpopulations it is only necessary to enter the analysis variable and classification variable. The program has the capability to screen for missing observations.

For regression equations, SUPER CARP can calculate coefficients, t-statistics, R-squares, and tests for groups of coefficients. Several options are available for equations with independent variables containing measurement errors: error variances known or estimated, reliabilities known or estimated, and error variances functionally related to the variable. Tests for the singularity of the matrix of true values can be calculated.

The package is designed primarily for variance estimation and analytic calculations, is rather restrictive on input format, and has little data management capability. SUPER CARP is written in FORTRAN. It is available on IBM computers.

E. Survey Analysis Programs

1. GENSTAT (General Statistical Program)

This package was developed by the Statistics Department at the Rothamsted Experimental Station in the United Kingdom. It is distributed by the Statistical Package Co-ordinator, NAG Central Office, 7 Bansbury Road, Oxford OX2 6NN, United Kingdom. GENSTAT provides a high level language for data manipulation and statistical analysis. It is used primarily for the analysis of experimental data, for fitting a wide range of linear and non-linear models, and for finding patterns in complex data sets using multivariate and cluster analysis. It can be used interactively or in batch mode. It is installed at approximately 150 sites.

Data can be presented in many different formats. Up to six way tables of totals, means, or counts can be formed and expanded to hold margins of means, totals, minima, maxima, variances, or medians. Tabular output in a wide range of layouts is provided.

Classical least squares regression, with or without weights, can be carried out. The fitting of linear models is generalized by providing functions linking the mean to the value predicted from the model and four error distributions. Non-linear models can be fitted using an iterative process. Designed experiments, including all balanced designs and many partially balanced designs, can be analyzed. Procedures for cluster analysis, principal component analysis, canonical variate analysis, and principal co-ordinate analysis are available. Time series analysis and forecasting are provided.

GENSTAT is written in FORTRAN. It is available for Burroughs, CDC, DEC, Honeywell, IBM ICL, PRIME, SIEMENS, UNIVAC, and VAX computers.

2. P-STAT

This package was developed and is distributed by P-STAT Inc. P.O. Box 285, Princeton, New Jersey 08540, U.S.A. P-STAT is a large conversational system offering flexible file maintenance and data display features, cross-tabulation, and numerous statistical

procedures. Its principle applications have been in areas such as demography, survey analysis, research, and education. The system can be used interactively or in batch mode. It is currently in use at over 100 installations around the world.

P-STAT provides for interactive data entry and editing. The edit file itself contains edit commands and data. It can be saved to be used again or submitted to run as a batch job.

In P-STAT many files can be active simultaneously. Commands are provided to update files, join files in either a left/right or an up/down direction, sort files on row labels or by up to 15 variables, and collate files which do not contain exactly the same cases, or which have a hierarchical relationship.

In batch mode, tables, frequency distributions, listings with labels, plots, and histograms may be easily produced. Once a table has been created, it can be modified conversationally without passing through the data again.

Chi-squares, F-tests, t-tests, means, and standard deviations are readily available. Commands are also provided to do correlations, regressions, principal components or iterative factor analysis, quartimax, varimax, or equimax rotations, and backwards-stepping multiple discriminant analysis.

P-STAT is written in FORTRAN and can be interfaced with SPSS and BMDP. It has been installed on Burroughs, CDC, DEC, Harris Honeywell, Hewlett-Packard, IBM, ICL, SIGMA 7, UNIVAC, and VAX computers.

3. FILAN

This package was developed and is distributed by ICL Dataskil, Reading, Berks, England. This is a system with a high level user language geared toward analysis of survey files of any size. The attributes of the command language require use by a high level language programmer familiar with survey requirements. There are three analysis programs in the package offering slightly different facilities. The operation of each of the survey analysis programmes can be discussed in terms of the following four phases:

(a) Data file creation phase. Validity checks for consistency of items can be done using a data dictionary. Error identification is provided for, but automatic correction is not possible. Variables can be recoded or transformed.

(b) Tabulation phase. A file of tables or matrices limited to two dimensions is created.

(c) Table manipulation phase. Mathematical manipulation of tables, matrices and table reorganization are possible.

(d) Output phase. Descriptive stubs and heading are added (ICL, 1980, pp. 51-56).

This system was developed in FORTRAN with ICL extensions and therefore is sold and supported by ICL for use on ICL computers. This system has in the past been used in certain UNSO-supported projects.

4. BIBLOS

This package was developed and is distributed by the French National Institute of Statistics and Economic Studies (INSEE), Paris, France. Called a "language for the statistician", this system is primarily a statistical analysis package. It provides a user command language in free-format coding and clear error messages for the syntactic checking of the user programme.

Phase 1 of this system includes file description, record selection, and definition of variables. New variables can be generated by computation. There is no theoretical limit to the size of the data files, as the system has a dynamic management of main storage and the files are never entirely brought into main storage. Any format of data which can be described in a FORTRAN FORMAT statement is acceptable.

Phase 2 is for data analysis and includes the following functions: principle components analysis; correspondence analysis; canonical analysis; hierarchical clustering; discriminant analysis; linear regression; segmentation; elementary statistical analysis; and dynamic clusters method.

Available documentation does not describe the package's ability to handle complex sample designs in the above analytical routines. Custom-coded FORTRAN routines can be inserted.

The system is available for IBM computers, having been written in a combination of FORTRAN, CPL/1, and Assembler. It requires a minimum of 192K bytes of primary storage. The user command languages is in French. Comprehensive documentation is available.

5. PACKAGE X

This package was commissioned by the United Kingdom Government Statistical Service and designed by ICL Dataskil (ICL's consultancy and software house). It is a general purpose system for direct use by statisticians with little knowledge of computers. It provides a powerful language facility packaged as a series of macro procedures. It has very rudimentary editing and correction facilities and provides the following general capabilities: summary statistics; significance testing; regression (multiple, stepwise, or polynominal); plotting; and tabulation.

This system has been described as an interactive dialogue retrieval system which closely questions the user as to his requirements. The parameters of the retrieval are extracted from the user's replies. PACKAGE X is specifically designed as a program-building system. Statisticians can construct their own specialized programmes from available building blocks.

Insufficient information is available to determine its computer requirements, beyond the fact that it is written in FORTRAN with ICL extensions, making its availability limited to ICL computers. No specifics are available to determine the robustness of its tabulations capabilities in relation to what is typically required for household survey tabular reports (ICL, 1980, pp. 50-51).

6. STATISTICAL ANALYSIS

This package was developed and is distributed by ICL Dataskil, Reading, Berks, England. This STATISTICAL ANALYSIS package is found in several national statistical offices and performs a variety of functions on input data that are assembled into an observational matrix, with or without missing values. The statistical measures include the following univariate and multivariate analysis routines: multiple regressions, canonical correlation; ANOVA; principle components; factor analysis; discriminant analysis; spectral analysis; and fourier analysis.

This system produces means, variances, weighted statistics, as well as transformed, normalized, cross-product, co-variance, and correlation matrices. Insufficient documentation as available to determine if the package can handle complex sample designs (ICL, 1980).

F. General Statistical Programs

1. BMDP

This package was developed and is distributed by BMDP Statistical Software, Department of Biomathematics, University of California, Los Angeles, California 90024, U.S.A. BMDP is a comprehensive library of general purpose statistical programs that are integrated by a common English-based control language and self-documented save files for data and results. Emphasis is placed on integrating graphical displays with analysis. The package is available for large and small computers. It is currently installed at over 1000 facilities throughout the world.

Data can be entered into BMDP from formatted files, binary files, free-formatted files, BMDP files, and a user definable subprogramme. Cases with illegal or implausible values can be identified. Various methods, including cross-tabulation, histograms, and bivariate scatter plots, are available for analysing all or part of a data file.

Two-way and multiway frequency table analyses include a wide variety or statistics. Linear regression features include simple, multiple, stepwise, all possible subsets, extensive residual analysis, detection of influential cases and multivariate outliers, principal component, stepwise polynomial, multivariate, and parital correlation. Non-linear regression includes derivative-based and derivative-free, stepwise logistic, and models defined by partial differential equations. Analysis of variance features include t-tests and one- and two-way designs with histograms and detailed statistics for each cell, factorial analysis of variance and co-variance including repeated measures, and balanced and unbalanced mixed models. Multivariate techniques include factor analysis, multivariate outlier detection, hierarchical clustering of variables and cases, k-means clustering of cases, partial and canonical correlation, multivariate analysis of variance, and discriminant analysis.

BMDP is written in FORTRAN. It has been installed on virtually every major computer.

2. SPSS (Statistical Package for the Social Sciences)

SPSS was developed and is distributed by SPSS, Inc., 444 N. Michigan Avenue, Chicago, Illinois 60611, U.S.A. SPSS is a computer package for the data analysis and file management. It is installed in over 2500 sites in 60 countries. The package runs on more than 30 different types of computers including IBM, ICL, CDC, Burroughs,

UNIVAC, Hewlett-Packard, DEC, and Honeywell, and is documented with a general manual, a primer, and an algorithm volume. Core requirements are 100-190K bytes depending on the version.

The package will accept input from cards, disk, or tape. It facilitates permanent or temporary data transformations, case selections, weighting, and random sampling of data. It allows for creation, updating, and archiving of system files containing a complete dictionary of labels, print formats, and missing data indicators. Up to 5000 variables can be defined for any one file. There is no built-in limitation to the number of cases. Provision is made for input and output of correlation, co-variance, and factor matrices. Other input includes z-scores, residuals, factor scores, canonical variates, and aggregated files. Report writing features include automatic formatting, a full range of summary statistics, composite functions across variables, and multiple level breakdowns. The following statistical analysis capabilities are provided:

(a) Frequency distributions, histograms, and descriptive statistics.

(b) Multiway cross-tabulations and measures of association for numeric or character data.

(c) Tabulation of multiple response date.

(d) Pearson, Spearman, and Kendall correlations.

(e) Partial correlation.

(f) Canonical correlation.

(g) Analysis of variance.

(h) Stepwise discriminant analysis.

(i) Multiple regression.

(j) Manova.

(k) Analysis of time series.

(l) Bivariate plots.

(m) Factor analysis.

(n) Guttman scale analysis.

(o) Non-parametric tests.

(p) Survival analysis.

(q) Pairs and independent samples t-test.

(r) Choice of treatment for missing values.

SPSS is written primarily in FORTRAN with a small amount of ASSEMBLER coding. A version of the package which does not require having a FORTRAN compiler is available.

3. OMNITAB-80

This package was written and is distributed by the National Bureau of Standards, Washington D.C. 20234, U.S.A. OMNITAB-80 is a high quality integrated general purpose programming language and statistical software computing system. The system enables the user to perform data, statistical, and numerical analysis with no prior knowledge of computers or computer languages. Simple instructions are used to reference varied and sophisticated algorithms for data analysis and manipulation. It may be used either interactively or in batch mode. OMNITAB-80 is transportable to any computer configuration sufficiently large to accommodate it.

OMNITAB-80 permits one to perform simple arithmetic, complex arithmetic, trigonometric calculations, data manipulation, special function calculations, statistical analysis, and operations on matrices and arrays. The system has extensive plotting, numerical analysis, and matrix analysis capabilities.

OMNITAB-80's statistical capabilities include one-way and two-way analysis of variance, regression, correlation analysis, cross-tabulation of any 14 statistics, contingency table analysis, and over 100 instructions for probability densities, cumulatives, percentiles, probability plots, and random samples. Almost all the statistical analysis instructions automatically provide comprehensive output.

OMNITAB-80 is written in FORTRAN. It has been installed on UNIVAC, IBM, CDC, Hewlett-Packard, and DEC computers.

4. SAS (Statistical Analysis System)

SAS was developed and is supported by SAS Institute Inc., Box 8000, SAS Circle, Carry, North Carolina 27511, U.S.A. It is a software system that provides tools for data analysis, including information storage and retrieval, data modification and programming, report writing, statistical analysis, and file handling. Since its beginning in 1966, SAS has been installed at over 2000 installations world-wide, and is used by statisticians, social scientists, medical researchers, and many others.

The SAS language is free-format with an English-like syntax. Data can be introduced into the system in any form from any device. Data management features include creating, storing, and retrieving data sets. SAS can handle complex files containing variable length and mixed record types, and hierarchical records. SAS has utility procedures for printing, sorting, ranking, and plotting data; copying files from input to output tapes; listing label information listing; and renaming, and deleting SAS and partitioned data sets. Report writing capabilities include automatic or custom-tailored reports with built-in or user-specified formats, value labels, and titles.

SAS offers 50 procedures for summary statistics; multiple linear or non-linear regression; analysis of variance and co-variance; multivariate analysis of variance; correlations; discriminant analysis; factor analysis; Guttman scaling; frequency and cross-tabulation tables; categorical data analysis; spectral analysis; autoregression; two- and three-stage least squares; t-tests; variance component estimation; and matrix manipulation.

SAS is written in PL/1 and IBM Assembler language. It includes a BMDP interface procedure, as well as a procedure for converting BMDP, OSIRIS, and SPSS system files to SAS data sets. SAS was designed orginally for IBM 360/37, and has been installed on Amdahl, Itel, National, Two Pi, Magnuson, Hitachi, and Nanodata computers. It requires a user region of 150K.

G. Data Management Programs

1. CENSPAC (Census Software Package)

This package was developed and is distributed by the Data User Services Division of the United States Bureau of the Census, Washington D.C. 20233, U.S.A. This newly released software is referred to as a generalized data retrieval system primarily for processing census public use statistical data files. It also has processing capabilities for summary data files and microdata files. These capabilities include: generalized input file definition; machine-readable data dictionaries; matching of two input files; sorting; record selection; report generation; extract file creation and documentation; interrecord and intrarecord computation and aggregation; array manipulation; and user subroutine and source code interface (United States Bureau of the Census, 1980, pp. 1-2).

Perhaps the more interesting aspects of the system are its abilities to match files, extract records, manipulate arrays and interface with user COBOL routines. Household survey operations such as questionnaire check-in and data linkage with other survey

rounds might be served by this package. This system does not have powerful language commands for comprehensive editing or tabulation of survey data.

This system is written in 1974 ANSI COBOL and requires 150K characters of primary storage and direct access storage. It presently is operational on IBM OS/VS and UNIVAC 1100 EXEC8 and can be converted to other vendor equipment, probably without difficulty as the system was designed to be machine independent.

The United States Bureau of the Census supports the package with limited training and software support in form of seminars and telephone and letter correspondence and is the clearing-house for modules developed by other users.

2. EASYTRIEVE

This package was developed by the Ribek Corporation in Naples, Florida, U.S.A. It is distributed by Pansophic Systems, Inc., 709 Enterprise Drive, Oak Brook, Illinois 60521, U.S.A. EASYTRIEVE is a system software tool for file maintenance, information retrieval, and report writing. It can be used by non-specialists, and can retrieve any kind of record from any file structure. It has over 2,000 users in more than 30 countries.

An EASYTRIEVE program, written in an English-like language, can use a few key words to call data and format a report. EASYTRIEVE programs can be run in a completely interactive mode using on-line systems, as well as in the batch mode; multiple jobs can be batched in a single program.

EASYTRIEVE provides for a wide variety of information retrieval. It can extract data from sequential, ISAM, VSAM, or data base files. It can access data of fixed, variable, undefined, or spanned record formats.

The package allows multiple input and output files. It supports creation and updating of files; matching and merging files; adding, deleting, and reformatting records; and providing audit trails for file updates.

Information analysis includes selection of data based on input, logic, and calculations; comparison of files; provision of conditional logic and calculation capabilities; performing table look ups and special tests; and sorting on up to 10 keys.

Multiple reports can be produced with one pass of the data. Reports are automatically formatted. Customizing alternatives to all report format features are provided. Summary reports and files can be produced.

EASYTRIEVE is written in IBM Assembler. It is available on IBM and IBM-compatible machines.

3. SIR

This package was developed and is distributed by SIR, Inc., P.O. Box 1404 Evanston, Illinois 60204, U.S.A. SIR is an integrated, research-oriented data base management system which supports hierarchical and network file structures.

It interfaces directly with SPSS and BMDP. It can be run either in batch or interactive mode. The program has been installed at over 70 sites world-wide. SIR is written in SIRTRAN, a macro preprocessor that generates FORTRAN and Assembler. The package is available on CDC, IBM, PRIME, SIEMENS, UNIVAC, and VAX computers.

A SIR data base is defined using SPSS-like data definition commands. These commands allow for multiple record types, the definition of hierarchical and network relationships, data editing and checking, data security at the item and record levels, and multiple data types. SIR provides a wide range of batch data entry and update options including new data only, replacement data only, and selected variable update.

The SIR retrieval language is structured and fully integrated with the rest of SIR. It has full arithmetic and logical operations. Retrieved information can be subjected to simple statistical analysis, used in reports, used to create SPSS or BMDP SAVE FILEs or a new SIR data base, or written to a formatted, external data set.

The interactive subsystem includes a text editor, storage of user-written procedures, and an interactive retrieval processor. The macro facility enables the creation of generalized procedures.

Other features include various utilities for restructuring, subsetting, merging, and listing, as well as automatic creation of journal files when data is added to or changed in the data base.

4. FIND- 2

This package was developed and is distributed by ICL Dataskil, Reading, Berks, England. The FIND-2 Multiple Inquiry System is labelled as a general purpose information retrieval and reporting package. The system allows the files to be interrogated based on specific criteria and provides for reorganization of files and record reformatting. It provides for custom program code to be easily inserted. Tabular analysis of data may be comprised of row or column totals, percentages, mathematical calculations, and summaries (ICL, 1980, pp. 47-49).

The system affords quick access to data but does not constitute a comprehensive package for complete processing of survey data files. It is available for ICL machines.

5. RAPID (Relational Access Processor for Integrated Data bases)

This system was developed and is distributed by the Special Resources Subdivision, Systems Development Division, R.H. Coats Building, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada KIA OT6. RAPID is a generalized data base management system which is typically used to process census and survey data. It is installed in six different statistical or government offices.

RAPID is based on the "relational" model, where all relations or files are viewed as simple matrices which contain rows (records) and columns (variables). A data base is thought of as any collection of RAPID files which are seen by a user as being related in some way. RAPID processes and manages data as well as data descriptions. The system provides access to this information by a consistent set of facilities which ensure integrity between the data and its description.

RAPID stores each relation in an IBM BDAM file as a fully transposed file which provides fast access for statistical retrievals. It uses its own data dictionary. A full set of data base administrator utilities are provided, including: utilities to create, expand, and shrink a RAPID file; backup and recovery programs; RAPID file analysis programs; and a single relation query facility.

Memory requirements vary depending on the physical characteristics of the RAPID files being processed. Most applications at Statistics Canada run between 200K and 512K bytes.

A RAPID-SPSS interface allows SPSS users to read RAPID files directly. New variables created during the SPSS job can be saved on the original RAPID file.

ANNEX II

MAJOR SOURCES OF TECHNICAL ASSISTANCE AND TRAINING IN DATA PROCESSING

In order to establish a capability to carry out a continuing programme of household surveys and to institutionalize skills in data processing, most countries participating in the NHSCP would require some form of technical assistance and training. There are several avenues available for obtaining this assistance, and some of the major ones are reviewed in this Annex although the coverage is by no means exhaustive. The information provided below is the most recent available at the time of writing and countries wishing to use these facilities will need to obtain the latest information from the institutions concerned. Each country will need to develop a plan which combines services provided by external sources with on-the-job, and local training in order to achieve the necessary capability.

Within the United Nations system, and more specifically within the context of the National Household Survey Capability Programme, technical assistance and training will be available at the national, regional and international levels. Where necessary, NHSCP country projects make provision for resident data processing advisers whose function is not only to assist countries in accomplishing the data processing task, but, more importantly, to train counterpart national staff and participate in in-house training programmes. At the regional level, the advisory services in data processing, include short-term consultancies, and organization of training seminars and workshops. In addition, several of the regional institutions that have been established with the support of the United Nations system offer programmes in data processing:

(a) The Statistical Institute for Asia and the Pacific (Tokyo) offers an introductory course in electronic data processing. From time to time, seminars are given in specialized fields, such as a seminar held on "Tabulation and Analytical Processing of Population Census Data" (United Nations, 1978).

(b) The Arab Institute for Training and Research in Statistics (Baghdad) offers an introductory course on computers and data processing (United Nations, 1978, p. 10).

(c) The Institute of Statistics and Applied Economics (Kampala) offers lectures on programming and data processing as part of the first year mathematics and statistics courses.

(d) The Institut National de Statistique et d'Economie Appliquee (Rabat) offers one year course to programmers and a three year program to analysts.

Institutions outside the United Nations system that offer data processing training include:

(e) The United States Bureau of Census (Washington, D.C.) offers technical assistance and training in data processing through its International Statistical Programs Center (ISPC). The one year course in computer data systems is designed to provide the knowledge and skills needed to qualify persons as systems analysts/ programmers, project managers, ADP (Automatic Data Processing) managers, and supervisors of computer operations, to train analysts to evaluate software and hardware; and to upgrade the capabilities of persons already specializing in computer data system. The training in systems analysis and programming languages is related primarily to third generation, medium-scale computers, of which the IBM System 360/370 series is representative. Participants are instructed in adapting languages and procedures to other types of equipment appropriate to the facilities in their own countries. Computer data systems is the major area of training emphasis, but several essential related fields are included, such as basic statistical concepts, design of tables and questionnaires, editing, coding, and imputation principles, and control and evaluation of non-sampling errors. A common request is for installation of one of the generalized software packages developed at the ISPC and training in its use. The ISPC has also developed a program of on-the-job training, whereby participants work with staff members in Washington to develop all or part of a system to process the data for a particular survey or census. The system is then installed in the participants' country and production running is monitored through short-term visits.

(f) The Center Europeen en Formation des Statisticiens-Economistes des Pays en Voie de Developpement (Paris) offers basic data processing skills, including the study of FORTRAN, are taught in both the first and second year classes.

(g) The International Statistical Education Centre (Calcutta) offers special courses on automatic data processing.

At the international level, support is provided by the United Nations Statistical Office through its interregional and technical advisory sevices, software development and dissemination activities, and general work on standards.

BIBLIOGRAPHY

Aldrich, Michael (1978a), "Data Entry Comes of Age." Data-Processing, Vol. 20, November, pp. 32-35.

Follows the history of data entry to present systems which are sophisticated computer systems in their own right and take data preparation beyond the punch room. Shows, through user application, how today's key-to-disk system can make a much greater contribution to an organization's efficiency.

_____(1978b), "Why Mainframes?" Data-Processing, July/August.

Follows the history of mainframe development and concludes that the mainframe is at the end of an era because of user demand for simplicity.

Allen, James (1977), "Some Testing and Maintenance Considerations in Package Design and Implementation." Interface, April, pp. 221-214.

Descriptions of design concepts which help programmers avoid errors and notes on procedures to follow to minimize errors during implementation and maintenance.

Alsbrooks, William T. and Foley, James D. (1977), "The Organization, Tabulation and Presentation of Data State of the Art: An Overview." Report on the Conference on Development of User-Oriented Software, November, pp. 63-65.

American Statistical Association (1977), Report on the Conference on Development of User-Oriented Software, November 1977.

A synopsis of papers, findings, and conference conclusions from sessions that sought the advice of experts outside the United States Bureau of the Census on research and development topics such mechanisms to improve access to and use of machine-readable census data; identification of software systems needed to assist the user community to more easily organize, tabulate and present census data; research and development activities that would lead to improvements and simplification to access and use of data; and recommendations to ASA on expansions to its programme.

Applebe, William and Volper, Dennis (1979), "A Portable Software Environment for Microcomputers." Interface, May, pp. 117-118.

Outlines the goals, design, and evolution of the University of California San Diego PASCAL System and how PASCAL provides microcomputer users with a software environment for problem solving and computer programming.

Australian Bureau of Statistics (1978, 1979), Statistical Computing Environments: A Survey, June 1978/revised February 1979.

Surveys the statistical computing facilities currently in use and planned for by the Australian Bureau of Statistics and other statistical agencies. Particular attention is paid to those systems which are sufficiently generalized and portable to be of direct use to the ABS. General topics such as data organization, programming languages, and computing hardware are covered as well as the development of integrated computing environments by five major agencies.

Banister, Judith (1980), "Use and Abuse of Census Editing and Imputation." Asian and Pacific Census Forum, Vol. 6, No. 3, February, p. 1. East-West Population Institute, Honolulu, Hawaii.

A thought-provoking article which thoroughly studies the pros and cons of editing and imputation and gives the reader a basis on which to make sound decisions in these areas of data processing.

Bessler, Joseph C. (1979), "OMR Systems Offer Time, Accuracy Benefits." Computerworld, Vol. 13 (June), p. 74.

Explores the advantages and disadvantages of OMR and discusses the situations for which OMR is most applicable.

Boehm, B.; Brown, J. and Lipow, M. (1976), "Quantitative Evaluation of Software Quality." Second International Conference on Software Engineering Proceedings, October, pp. 592-605.

A report of a study done by TRW Systems and Energy Group that establishes a conceptual framework and some key initial results in the analysis of the characteristics of software quality.

Brophy, Hugh F. (1977), "Generalized Statistical Tabulation." Report on the Conference on Development of User-Oriented Software, November, pp. 207-208.

Chambers, John (1979), "Designing Statistical Software for New Computers." Interface, May, p. 100.

_____(1980), "Statistical Computing: History and Trends." The American Statistician, November, pp. 238-243.

Looks at history and current trends in both general computing and statistical computing, with the goal of identifying key features and requirements for the near future. Also includes a discussion of the S language developed by Bell Laboratories.

Cottrell, Samuel IV and Fertig, Robert T. (1978), "Applications Software Trends: Evolution or Revolution?" Government Data Systems, Vol. 7, January/February, pp. 12-13+.

Looks at the problem of present day software development and explores ways to bring software costs back in line with hardware costs.

Datapro Research Corp. (1978), "Build or Buy Software? That is the question." Computerworld, Vol. 12, September, pp. S16 onwards.

Poses a list of questions to consider in making a decision on whether to build or buy software.

Delta Systems Consultants, Inc. (1979), Report on Computer Hardware Available in Developing Countries for Processing Census Data, December.

An inventory of data entry systems, mainframe computer systems and mini computer systems available for use by statistical offices, in program countries of the United States Agency for International Development.

Durniak, Anthony (1978), "Computers." Electronics, October, pp. 152-161.

Discusses advances in hardware technology and their impact on the computer industry.

Fellegi, I.P. and Holt, D. (1976), "A Systematic Approach to Automatic Edit and Imputation." Journal of the American Statistical Association, Vol. 71, No. 353, pp. 17-35.

Presents an in-depth discussion of the Fellegi-Holt technique of editing and imputing data. Provides supporting mathematical theory.

Ferber, Robert; Sheatsley, Paul; Turner, Anthony and Waksberg, Joseph (1981), What is a Survey? Washington D.C.: American Statistical Association.

Describes survey operations without using technical terminology; understandable by persons not trained in statistics.

Francis, I. and Sedrank, J. (1976), "Software requirements for analysis of surveys." Proceedings of the Nineth International Biometric Conference, pp. 228-253.

Francis, Ivor (1979), A Comparative Review of Statitical Software.

A very comprehensive description and critique of 46 packages available for statistical computing. Responses to a questionnaire sent to the developers are tabulated.

Frank, Werner L. (1979), "The New Software Economic." Computerworld, Vol. 13, January 15, 22, 29 and February 5.

Surveys the software life cycle and the productivity issue; discusses the status of the software products industry, identifies successful software products, and probes the criteria which must be satisfied for success; focuses on the software product supplier, developing financial models that contrast the software supplier's economics with those of the hardware manufacturer; and summarizes the problems and promise of the software products industry.

_____(1980), "Software Maintenance Here to Stay." Computerworld, November 24, pp. 35, 38.

Points out reasons why the maintenance effort in software will continue to grow and offers suggestions for containment of the effort required.

Friedman, Herman (1979), "The Use of Graphics Software in Concert with Multivariate Statistical Tools for Interactive Data Analysis." Interface, May, pp. 160-168.

Touches on key aspects of hardware, systems support, languages, and application packages necessary for graphics as part of interactive statistical analysis.

Gantt, M.D. (1979), "First Buyers: Beware of Great Expectations." Computerworld, Vol. 13 (January), p. S/4.

Hetzel, William C. and Nancy L. (1977), "The Future of Quality Software." IEEE Computer Society Conference Proceedings, Spring, pp. 211-212.

Surveys trends in software development and traces the impact on the quality of software that is produced.

Hill, Gary L. (1977), "The Generative Approach to Software Development." Proceedings, ACM National Conference, pp.68-73.

Describes the generative programming techniques employed by the CENTS-AID II system.

Hill, Mary Ann (1977), "Current BMDP Research and Development." Interface, May, pp. 376-378.

Results of two preliminary programs (BMDQ1T, BMDQ2T) for time series analysis are described and their capabilities are displayed.

Hursch-César, Gerald and Roy, Prodipto (1976), Third World Surveys: Survey Research in Developing Nations. New Delhi: The Macmillan Company of India, Limited.

Attempts to show where major and sometimes drastic improvements are needed in survey design, conduct, and interpretation. Focuses on some of the common practical and intellectual problems faced by investigators when they engage in survey research in developing countries.

Institute National de la Statistique et des Etudes Economiques (1975), LEDA: Statistician's Manual, second edition, p. 1.

Kaplan, Bruce; Francis, Ivor and Sedransk, J. (1979), "Criteria for Comparing Programs for Computing Variances of Estimators from Complex Sample Surveys." Interface, May, pp. 390-395.

Criteria for comparing computer programs for calculating variance estimators of point estimators from complex sample surveys are presented along with various methods of estimating variances including Taylor series expansion, balanced repeated replications, jackknifing and the Keyfitz method. Three packages are described (CLUSTERS, STANDARD ERROR, and SUPER CARP) and their performance is measured.

Kaufman, Felix (1978), "Distributed Processing." Data Base, Vol. 10 (Summer), pp. 9-13.

Discusses the evolution of distributed processing and concludes that it is the direction of the foreseeable future.

Khoo, Siew-Ean; Suharto, Sam; Tom, Judith A. and Supraptilah, Bondan (1980), "Linking Data Set: The Case of Indonesia Intercensal Population Survey." Asia and Pacific Census Forum, Vol. 7, No. 2, November. East West Population Institute, East West Center, Honolulu, Hawaii.

A detailed discussion of the procedure followed in linking data from an Indonesian survey of fertility to two other data sources. A good example of the successful use of exact matching.

Krug, Doris (1980), Costs of Four Recent Household Surveys. WESTAT memorandum to B. Diskin.

Detailed cost information for four household surveys conducted in the United States by WESTAT.

Lackner, Michael and Shigematsu, Toshio (1977), "Some Statistical Data Processing Software for Small Computers." Bulletin of the International Statistical Institute, Vol. 47, Book 1, pp. 265-276.

A description of the philosophies behind and the function and capabilities of the generalized software developed by the United Nations Statistical Office (XTALLY, UNEDIT).

Lusa, John M. (1979), "Going to the Source." Infosystems, Vol. 26 (April), pp. 52-56.

Traces the evolution of source data entry and examines future trends in data entry.

Muller, M.E. (1980), "Aspects of Statistical Computing: What Packages for the 1980's ought to do." The American Statistician, Vol. 34, No. 3.

Nelson, David (1979), "An SPSS-Compatible Software Exchange - Plans and Goals." Interface, May, pp. 424-426.

Description of the definition of the goals of SPSS software exchange, soliciting quality routines for the exchange, preparing and distributing catalogues of available software, arranging conversions of selected programs to various hardware installations, and distributing user documentation.

Nelson, Tolbert and Soper (1979), "An SPSS-Compatible Software Exchange Plans and Goals." Interface, May. pp. 424-425.

Newcombe, H.B.; Kennedy, J.M.; Axford, S.J. and James, A.P. (1959), "Automatic Linkage of Vital Records." Science, October, pp. 1-6.

Describes one of the earliest attempts at computerized data linkage of marriage and birth records from the Canadian province of British Columbia.

Okner, Benjamin A. (1972), "Constructing a New Data Base from Existing Microdata Sets: the 1966 Merge File." Annals of Economic and Social Measurement, Vol. 1, No. 3.

Contains a detailed explanation of the procedures used to construct the 1966 MERGE file, a microdata source which contains information from the 1967 Survey of Economic Opportunity and the 1966 Tax File (United States).

_____(1974), "Data Matching and Merging: An Overview." Annals of Economic and Social Measurement, Vol. 3, No. 2, p. 348.

Summarizes discussion at a data matching and merging workshop. Gives extensive background information on the theory and application of data linkage.

Rattenbury, Judith (1980), "Survey Data Processing - Expectations and Reality." Paper presented at the World Fertility Survey Conference, London, July 7-11.

Presents a realistic look at many of the problems associated with survey data processing and practical ideas for confronting them. Based on experience gained through the World Fertility Survey program.

Rhodes, Wayne L., Jr. (1980), "The Disproportionate Cost of Data Entry." Infosystems, October, pp. 70-76.

Assesses the current state of data entry and discusses ways to improve quality and reduce cost.

Ross, Ronald B. (1978), "Data Base Systems: Design, Implementation and Management." Computerworld, May 22, 29 and June 5.

A very comprehensive look at data base systems which would provide the potential user with a good background for understanding what he or she is undertaking.

Rowe, B. (1980a), "Outline of a Programme to Evaluate Software for Statistical Processing." Statistical Software Newsletter, Band 6, Heft 1, pp. 5-8.

Rowe, Beverley (1980b), Statistical Software for Survey Research. World Fertility Survey, London.

A listing of approximately 100 statistical packages indicating function, host language, compatible hardware, and distribution source.

Ruggles, Nancy; Ruggles, Richard and Wolff, Edward (1977), "Merging Microdata: Rationale, Practice and Testing." Annals of Economic and Social Measurement, Vol. 6, No. 4, pp. 416-417.

A three-part paper which argues for the need for the statistical matching of microdata sets as a way of reconciling diverse bodies of data, discusses one particular matching technique developed at the National Bureau of Economic Research, and performs several econometric tests to evaluate the reliability of the matching technique.

Sadowsky, George (1977), "Report on a Mission to Bolivia." United Nations, 1977 (photocopied).

Discusses the interaction of the National Statistics Office (INE) with a central computer centre (CENACO) and presents many of the problems inherent in this relationship.

_____(1978), "Report on a Mission to Bolivia." United Nations, 1978 (photocopied).

Updates an earlier report and proposes obtaining a minicomputer for the National Statistics Office (INE).

_____(1979a), "Report on a Mission to the Republic of Cape Verde." United Nations, 1979 (photocopied).

Discusses the actual installation of minicomputers for the purpose of procesing a census.

_____(1979b), "Report on a Mission to Suriname." United Nations, 1979 (photocopied).

Presents a good overview of software available for editing and tabulation and a computerized approach to input/output control.

_____(1980), "Report of a Mission to Botswana." United Nations, 1980 (photocopied).

Presents ideas for processing a continuous household survey programme in Botswana called CHIPS, under auspices of the National Household Survey Capability Programme.

Scott, Christopher (1973), Technical Problems of Multiround Demographic Surveys. Chapel Hill, North Carolina: Laboratories for Population Statistics.

A discussion of some of the practical problems associated with follow-up surveys in the context of their use in developing countries, giving specific recommendations wherever possible.

Smith, Martha E. and Newcombe, H.B. (1975), "Methods for Computer Linkage of Hospital Admission-Separation Records into Cumulative Health Histories." Methods of Information in Medicine, July, pp. 118-125.

Description of a study involving the design and testing of a computer system for linking hospital admission-separation records into longitudinal health histories.

Swanson, E.B. (1976), "The Dimension of Maintenance." Proceedings of the Second International Conference on Software Engineering, October, p. 494.

Taylor, Alan (1980), "Independent COBOLS Fit New Software Pattern." Computerworld, Vol. 14, pp. 31-32.

Discussion of vendors' attempts to produce COBOL compilers containing pseudocode structure that would be independent of hardware and operating systems.

United Nations (1964), Department of Economic and Social Affairs, Statistical Office Recommendations for the Preparation of Sample Survey Reports. Statistical Papers, Series C. Rev. 2, Sales No. 64.XVII.7, p. 3.

Provides recommendations for the preparation of preliminary, general and technical reports on surveys, and defines some key survey sampling terms.

_____(1977), The Organization of National Statistical Services: A Review of Major Issues. Studies in Methods, Series F., No. 21, Sales No. E.77.XVII.5.

Presents a review of major issues associated with the organization of national statistical services.

_____(1979a), Studies in Integration of Social Statistics: Technical Report and Methods. Studies in Methods, Series F, No. 24, Sales No. E.79.XVII.4.

A technical report encompassing studies in the integration of social statistics.

_____(1979b), Improving Social Statistics in Developing Countries: Conceptual Framework and Methods. Studies in Methods, Series F, No. 25, Sales No. E.79.XVII.12.

Presents a conceptual framework for improving social statistics in developing countries.

_____(1980a), National Household Survey Capability Programme: Prospectus, DP/UN/INT-79-020/1.

Describes the nature and purpose of the NHSCP, its organization feature, scope and requirements.

_____(1980b), Draft Handbook of Household Surveys, DP/UN/INT-79-020/2.

Forms the basic document for NHSCP technical studies of which the present study is one. Published in four volumes covering general survey planning; issues in survey content, design and operations by substantive area including demography, income and expenditure, employment, food consumption and nutrition, agriculture, health, education and literacy; selected issues from regional survey experience; and examples of survey questionnaires used in countries.

_____(1980c), Handbook of Statistical Organization, Vol. I. Studies in Methods, Series F. No. 28, Sales No. E.79.XVII.17.

A study of the organization of National Statistical Services and related management issues.

_____(1978a), Economic and Social Council. Review of Training of Statistical Personnel, E/CN.3/525, April 17.

Reviews the training of statistical personnel carried out through the United Nations system at the regional level and by selected international and regional institutions outside the United Nations system.

_____(1978b), ESCAP Computer Information. Economic and Social Commission for Asia and the Pacific, A.D./39, November 1978.

_____(1980d), Report of African Statistical Data Processing 1978-1979. Economic Commission for Africa, January 1980.

An in-depth report by the Conference of African Statisticians on responses from 145 organizational units in 50 countries from the fourth survey on data processing capabilities and requirements conducted in 1978. It includes an inventory of electronic data processing equipment, related staff resources, and applications for both the private and public sectors.

_____(1980e), Progress Report on Statistical Data Processing, E/CN.3/535, July 1980.

Provides a general description of the coverage of the United Nations technical co-operation activities in data processing, a description of the United Nations Statistical Office programme for the development of computer software, and a description of present day data processing hardware.

United States Department of Commerce, Bureau of the Census (1974), UNIMATCH 1 Users Manual: A Record Linkage System.

A users guide to a generalized record linkage system which permits the user to define a matching algorithm suitable.

_____(1979), Popstan: A Case Study for the 1980 Censuses of Population and Housing, Parts A and B.

A comprehensive study of all aspects of carrying out a census of population and housing in the mythical country of Popstan. An invaluable resource for developing countries attempting to take a national census.

_____(1980), Developing World Computer Facts.

A compilation of facts and figures concerning available computer hardware and software packages at government-run or government-used installations in developing countries world-wide. Includes individual country enumeration of existing computer facilities. Data sources include United States Bureau of the Census files, United Nations/ESCAP reports, IASI reports and FAO reports.

United States Department of Commerce, National Bureau of Standards (1977), Accessing Individual Records from Personal Data Files using Non-Unique Identifiers.

Presents selected methodologies for assisting federal agencies in selecting retrieval algorithms and name look up techniques; in analyzing their data by the identification of weighting factors and statistical sampling for determining error and omission rates; and in predicting the accuracy and efficiency of candidate retrieval keys.

United States Department of Commerce, Office of Federal Statistical Policy and Standards (1980), Report on Exact and Statistical Matching Techniques.

Describes and contrasts exact and statistical matching techniques. Discusses applications of both exact and statistical matches. Intended to be useful to statisticians in determining which technique is appropriate to a situation.

United States Department of Health and Human Services, Social Security Administration (1980a), Report No. 3: Matching Administrative and Survey Information: Procedures and Results of the 1963 Pilot Link Study

Describes methods employed in the 1963 Pilot Linkage Study to search for income tax and social security records. Primary focus is an examination of reporting differences between survey and administrative sources.

_____(1980b), Report No. 11: Measuring the Impact on Family and Personal Income Statistics of Reporting Differences Between the Current Population Survey and Administrative Sources.

A collection of papers examining income reporting differences between the Current Population Survey (CPS) and Social Security or Federal income tax records. Most of the results taken from the 1973 Exact Match Study.

Verma, Vijay and Pearce, M.C. (1978), Users' Manual for CLUSTERS. London: World Fertility Survey.

Verma, V.; Scott, C and O'Muircheartaigh, C. (1980), "Sample Designs and Sampling Errors for the World Fertility Survey." Journal of the Royal Statistical Society, Vol. 143, Part 4, pp. 431-473.

Wagner, Frank V. (1976), "Is Decentralization Inevitable?" Datamation, November, pp. 86-97.

Asserts that the repeal of Grosch's Law by technical advances makes a clear case for decentralization of computing. Sets forth "the principle of decentralized computing."

Weinberg, J. and Yourdan, E. (1977), "State of the Future." Trends in Software Science, Vol. 28, June, p. 39.

Article abstracted from a seminar discussion which emphasizes responsive design of software systems.

Wiederhold, Gio (1977), Database Design. New York: McGraw Hill Book Company

Presents the methods, the criteria for choices between alternatives, and the principles and concepts that are relevant to the practice of data base software design.

Wilkinson, G.N. (1977), "Language requirements and designs to aid analysis and statistical computing." Bulletin of the International Statistical Institute, Vol. 47, Book 1, pp. 299-311.

Wiseman, Toni (1977), "Questions Urged on Users Debating Software Options." Computerworld, Vol. 11, April, p. 18.

Poses questions that should be answered in the course of whether to go outside for software development as opposed to having one's staff develop the needed software.

World Fertility Survey, International Statistical Institute, London (1976), Editing and Coding Manual. Basic Documentation No. 7.

Provides useful guidelines on planning and designing of manual editing and coding operations, specifically for WFS surveys.

_____(1980), Data Processing Guidelines. Basic Documentation No. 11.

One of the outstanding documents describing procedures for specification, implementation and documentation of data processing for a survey. Written largely in the specific context of WFS surveys.

Yaskai, Edward K. (1978), "Wanted: More Power." Datamation, April, pp. 187-188.

Presents Seymour Cray's argument for further development of large scientific computers to provide computing power thousands of times more available than anything now available.

Yates, Frank (1975), "The Design of Computer Programs for Survey Analysis: A contrast between 'The Rothamstead General Survey Package' (RGSP) and SPSS." Biometrics, 31, pp. 573-584.

Zelkowitz, Marvin (1979), "Resource Estimation for Medium-scale Software Products." Interface, May, pp. 267-272.

Describes the Software Engineering Laboratory at the University of Maryland and NASA Goddard Space Flight Center for studying the mechanics of medium-scale development.

Printed in U.S.A.

40734-July 1991-500