



STUDIES IN METHODS

Series **F** No. **19**

**ORGANIZATION
AND CONDUCT
OF DISTRIBUTIVE-TRADE
SURVEYS**

UNITED NATIONS

DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS
STATISTICAL OFFICE

STUDIES IN METHODS

Series **F** No. **19**

**ORGANIZATION
AND CONDUCT
OF DISTRIBUTIVE-TRADE
SURVEYS**



UNITED NATIONS
New York, 1977

NOTE

Symbols of United Nations documents are composed of capital letters combined with figures. Mention of such a symbol indicates a reference to a United Nations document.

ST/ESA/STAT/SER.F/19

UNITED NATIONS PUBLICATION

Sales No. E.77.XVII.3

Price: \$U.S. 10.00
(or equivalent in other currencies)

CONTENTS

	<u>Page</u>
INTRODUCTION	
General remarks	1
Organization of the manual	2
<u>Chapter</u>	
I. PRELIMINARY STEPS AND MAJOR CONSIDERATIONS	
Study of existing programmes	4
Study of international recommendations	5
Major considerations	6
II. NEEDS AND USES OF DISTRIBUTIVE-TRADE STATISTICS	
General remarks	11
Governmental needs and uses	11
Business needs and uses	12
Other needs and uses	14
III. GENERAL PLANNING AND ORGANIZATION	
Preparing a brief for approval of the programme	16
Legislative requirements	16
Determining organizational structure	17
Assembling the planning staff	18
Organizing advisory groups	19
Fundamental decisions	21
Method of assembling data	22
Method of data processing	22
Equipment and staff requirements	23
Allocation of responsibilities	23

	<u>Page</u>
Co-ordination with other surveys	24
Planning a pre-test	26
Publication planning	27
Differences between planning for census programmes and planning for survey programmes	28
IV. MANAGEMENT AND BUDGET	
Use of manuals	30
Use of flow charts	31
Time schedules	31
Scheduling staff and equipment use	33
Progress reporting system	33
Establishing a quality control system	35
Maximizing the use of resources	36
Differences between censuses and current surveys	38
V. SCOPE, COVERAGE AND CLASSIFICATION	
Introduction	40
Scope	41
Coverage	43
Classification	48
VI. QUESTIONNAIRES AND OTHER DATA-COLLECTION FORMS	
Information needed for control, correspondence and imputation	49
Questions for publication and editing	50
Differentiating non-reporting entries from zero entries . .	51
Designing the kind-of-activity question	52
Detecting consolidated reporting	53
Keying the report form	53
Importance of pre-coding	54
Treatment of instructions	54

	<u>Page</u>
Problems arising from the diversity of activities	55
Design of report forms	57
Sequence and phrasing of questions	58
 VII. DATA COLLECTION	
General remarks	61
Methods of enumeration	63
Field enumeration	64
Establishing and using a directory	68
Collection and control of reports	76
Use of sampling	81
 VIII. DATA PROCESSING	
General remarks	86
Clerical operations	87
Preparation for tabulation of the questionnaire	90
Computerized editing and coding	94
Data from administrative records	98
Editing in repetitive sample surveys	99
 IX. TABULATION, REVIEW AND CORRECTION	
General remarks	101
Adding machines	101
Electro-mechanical processing	102
Computer processing	105
Review and correction of tabulations	106
 X. PUBLICATION	
Final review	112
Final disclosure analysis	112
Copy preparation and duplication	113

	<u>Page</u>
Publication of preliminary reports	114
Inquiry descriptions to be published	115
Content of tables	116
More-frequent-than-annual surveys	122

ANNEXES

I. SAMPLING TECHNIQUES IN SURVEYS OF DISTRIBUTIVE TRADES AND SERVICES .	124
II. BIBLIOGRAPHY	164

INTRODUCTION

General remarks

1. The Statistical Commission of the United Nations at its eighteenth session in 1974, inter alia, adopted recommendations on statistics of the distributive trades and services, endorsed the proposal of the Secretary-General and recommended "the preparation of a manual on the organization and conduct of surveys, to be issued as a methodological supplement to the international recommendations". ^{1/}
2. This publication has been prepared in response to that recommendation and is intended as a practical guide to the planning and management of censuses and surveys of distributive trades and services, with particular attention to the problems faced by countries at an early stage of statistical development.
3. The conceptual framework of the inquiries for which the practical aspects of their undertaking are discussed in this manual is outlined in the international recommendations adopted by the Statistical Commission. ^{2/}
4. In view of the similarities between industrial and distributive-trade inquiries, this manual incorporates a discussion of some methods and procedures that are also applicable to industrial surveys. However, since at present most countries conduct industrial and distributive trade inquiries separately, it is thought that separate manuals, by providing more specialized guidance, would be easier to use by those countries planning to conduct such inquiries.

^{1/} Official Records of the Economic and Social Council Fifty-eighth Session, Supplement No. 2 (E/5603), para. 97.

^{2/} International Recommendations on Statistics of the Distributive Trades and Services, Statistical Papers, Series M, No. 57 (United Nations publication, Sales No. E.75.XVII.9).

5. For the purposes of the present manual, a census is conceived as a comprehensive inquiry conducted at infrequent intervals, usually of 5 or 10 years. Current surveys are those conducted at frequent intervals - monthly, quarterly or annually. Censuses and surveys encompass the activities of wholesale and retail trade, restaurants and hotels and selected services as defined in International Standard Industrial Classification of All Economic Activities (ISIC) ^{3/} major division 6 and parts of divisions 83, 94 and 95. The units for which statistics are compiled are business establishments (stores, shops, offices or other single locations) within the territorial boundaries of the country. Enterprises are not the basic statistical units, although in some cases they may be the reporting units.

Organization of the manual

6. The present manual attempts to discuss the steps that should be taken in order to plan and conduct a distributive-trade inquiry. Its content has been organized, whenever possible, following the sequence in which the operations should be planned for the inquiry. In many aspects, the methods used for infrequent censuses and current surveys are quite similar. However, when a different treatment is required, this has been so indicated.

7. Chapter I describes the preliminary steps and major considerations for the organization of the inquiries, chapter II describes the needs and uses of distributive-trade data, chapters III and IV are devoted to discussions of general planning and organization, management and budget. Chapters V and VI examine questions related to the scope, coverage, classification and design of forms and

^{3/} International Standard Industrial Classification of All Economic Activities, Statistical Papers, Series M, No. 4, Rev. 2 (United Nations publication, Sales No. E.68.XVII.8).

questionnaires. Chapter VII considers methods and problems of data collection. Chapters VIII and IX describe the steps and problems associated with data processing, tabulation, review and correction. The manual ends with chapter X, a discussion of the publication of results. A detailed discussion on practical sampling techniques in distributive-trade inquiries is provided in annex I.

8. The manual should be regarded only as a set of practical suggestions on the most important aspects of organizing and carrying out distributive-trade inquiries. In the brief compass of this document it has not been possible to discuss every subject in adequate detail and the recommendations included merely constitute a guide to certain aspects of taking a survey. Every country contemplating such an undertaking will need to study these recommendations further in order to adapt them to its particular conditions.

9. Although international recommendations are intended to reflect common practices in the majority of countries, those interested in a more detailed discussion of certain topics or in the particular experience of certain countries should communicate with such countries to determine the problems they might have encountered and to benefit from their solutions. Advice on particular problems may also be obtained by writing to the Director, Statistical Office, United Nations, New York, N.Y. 10017.

I. PRELIMINARY STEPS AND MAJOR CONSIDERATIONS

Study of existing programmes

10. While the international recommendations have suggested the kinds of data to be collected in the field of distributive trades and services, it is necessary for each country to review for itself the data already being collected and what its objectives should be. This requires the study of administrative reports being collected by various government agencies and a thorough investigation into the data needs of private and public sectors.
11. A review of government functions may reveal that a wealth of data is already available in connexion with the operation of taxing agencies, the regulation of financial organizations, the administration of labour laws or the operation of social welfare activities. These data need to be examined to determine whether they are usable in statistical programmes in addition to the regulatory or administrative functions that required them originally.
12. To determine data needs, it will be necessary to contact two major classes of users, governmental and private. It is worth noting that decision- and policy-making units of the Government frequently determine their actions without adequate statistical information. Were they given the opportunity of voicing their needs, their requests for data could well be overwhelming. The problem would then be one of limiting the programme to feasible levels.
13. One way of exploring the data needs of private users would be to see what data programmes private organizations undertake. It will usually be found that such programmes are statistically unreliable because of their inability to achieve high and representative response rates, but they are conducted because no other approach is possible. The fact that the organization is, nevertheless, driven to make the effort is a good indication of the usefulness of the data.

14. In reviewing data needs, it should be kept in mind that there are two major, albeit highly integrated, types of programmes that need to be considered. The first consists of a comprehensive census which attempts to develop statistics in comparatively great detail as a basis for understanding the structure of the entire sector. Such a census is usually taken at infrequent intervals, perhaps 5 or 10 years apart, and requires a long period of time for planning and processing. The second type of programme is conducted at more frequent intervals - monthly, quarterly, annually, is much more restricted as to subject matter and may be conducted on a sample basis. While the samples used in current surveys do not necessarily have to be selected from complete census enumerations, they will be more reliable and useful if so based. This permits tying the two sets of data together and acts as a check on the functioning of the samples between census periods.

Study of international recommendations

15. As indicated above, the international recommendations adopted by the Statistical Commission and since published by the United Nations ^{4/} are available to assist countries in developing their distributive-trade statistics. The publication includes discussions on the following subjects: the scope, coverage and units used; the characteristics of the statistical unit and its classification; the items of data to be gathered and statistics to be published; and the definitions of items of data. The programme outlined includes recommendations concerning infrequent inquiries (censuses), annual inquiries and more-frequent-than-annual inquiries. For each category, recommendations are made concerning which items of

^{4/} International Recommendations on Statistics of the Distributive Trades and Services, Statistical Papers, Series M, No. 57 (United Nations publication, Sales No. E.75.XVII.9).

data should be included. Discussions of major problem areas, especially the definitions of statistical and reporting units and definitions of establishment characteristics, such as kind of activity, type of operation, kind of economic organization, kind of legal organization and type of ownership, location and size, are also provided. It is not expected that these recommendations will be found acceptable in every detail by all countries. Appropriate modifications should be made dependent on each country's special circumstances. An attempt should be made, however, to fit the national programme into the framework of the international recommendations so that meaningful comparisons can be made from country to country and world statistics can be further developed on the basis of standard concepts. As the economics of the world become more integrated and interdependent, standardization can lead to a better understanding of the interplay of economic forces and a more intelligent solution of economic problems.

Major considerations

16. Because of the great complexity of both the infrequent censuses and the current surveys, it is important that all aspects of the work to be performed be understood and planned for. It is necessary to consider carefully each operation and its interactive effect to make sure that the final product will be acceptable. This is especially true for infrequent censuses for which it will take 5 to 10 years to make the necessary modifications in the specifications. For the more frequent surveys, corrections can be made more frequently and the effect of an error will not be so disastrous. The development of a well thought out programme specifically adapted to the needs of the particular country is important. Statistical systems that work well in one country may be quite inappropriate for use in another country.

17. A proper and realistic budget is a function of thorough advance planning and

understanding cost allocation. Frequently, the person responsible for the technical aspects of the programme is also burdened with the responsibility of budgeting, with the result that costs are underestimated. It is important that a competent budget analyst review all operations to make sure that the budget is proper and sufficient. Contingency reserves must be provided for in the event of emergencies. The operation must be thoroughly analysed to make sure nothing of significance is omitted. Because of poor planning and budgeting, some countries have found that their funds were largely expended in the collection phase and no resources were left to complete the processing of questionnaires and publication of results.

18. A very common shortcoming, especially in infrequent censuses, results from overloading the programme and complicating the questionnaire to the point that the entire system breaks down because of its complexity. While it is important to include vital information, such information may be lost either because the respondent balks at the great volume of data requested or does not have some of the data readily available, or because it becomes impossible to process the data because they are so elaborate. It is especially important in initiating a new programme to keep the content of the questionnaire as simple as possible to assure the successful conclusion of the census.

19. To have a successful survey, it is essential to obtain the co-operation of the respondents. One way of obtaining such co-operation is through legal compulsion, and the provision of penalties for non-response is probably a necessity in the case of comprehensive censuses. The case for current surveys based on samples is not so strong, since in such surveys the respondent selected in the sample is asked to assume a burden not universally applied. In any

event, a willing collaborator is more apt to supply a more correct and timely report than one who is recalcitrant. Therefore, it behooves the survey manager to work with business organizations and representatives of large firms to establish a proper climate for the conduct of the survey. Sometimes this means the inclusion of some special item of data wanted by a specific industry and sometimes the exclusion of an unwanted item. It also means that the data requested must be readily obtainable from the business records of the respondents. A rather frequent form of complaint by business firms is that the requested data have already been supplied to another governmental agency. While this may be true, it is often necessary to request the same information again because of timing or the complexity of transferring the data from the other agency.

20. As mentioned above, it may be possible to use regulatory agencies for the purposes of the inquiry. First, a mailing list is usually available or could be developed from their records. Such a list frequently contains addresses in addition to industry classification and perhaps some kind of data such as volume of sales. This information can be very helpful in developing a directory for use in the census or in the preparation of samples for current surveys. Secondly, if the regulatory agency is already collecting data, it may be possible to ask it to modify its requests in order to accommodate the needs of the inquiry.

21. Use of the records of other governmental agencies will frequently require skilful negotiation, since such agencies are not primarily concerned with statistics. It will also require careful planning to make sure that such records become available on time and in the proper form for use in the survey. A problem may arise because data from a regulatory agency may have been collected under a promise of confidentiality. Since the surveys will also be collecting data under a law that

promises confidentiality, the use of the data from the regulatory agency can be justified so long as the statistical agency personnel are sworn to confidentiality and the publications contain no data that will disclose the operations of any firm. The converse, however, is not true; that is, information on individual firms collected in the statistical survey cannot be transferred to a regulatory agency, since such information was obtained with the promise that it could not be used for regulatory or taxing purposes.

22. A number of items of data in surveys of distributive trades and services are similar to data items contained in other surveys being conducted either by the statistical agency or by other governmental units. It is important to ensure, as much as possible, that common items are defined in the same way, so that data from different surveys may be compared or aggregated. Modifications that may be required because of special conditions in a particular industry should be designed to make the resulting data comparable. This is especially important with respect to the same data being collected in a comprehensive census and in a current survey. For example, because of the necessity for the rapid collection and publication of data in a monthly survey, the respondent may have some difficulty in adjusting his sales data for such items as value-added taxes, trade-ins, refunds etc. This could lead to non-comparability of data between censuses and current surveys.

23. In most countries, there is usually a data-collection organization already in existence. A decision has to be made concerning the extent to which the different surveys can or should be integrated, since there are usually a number of common elements. While it is possible to integrate many of the processing operations, the planning functions, especially as they relate to data content and industry knowledge, need to be provided for in a separate organization. Even in the case of the processing operations, care should be taken not to lose sight of the fact that knowledge of specific industries is important in assessing the nature of problems and how they should be solved.

24. Many surveys turn out to be unsuccessful despite great efforts made prior to conducting them, because the final test can come only from the actual conduct of the survey. In all cases, the minimum activity prior to launching a survey should include visits to business firms to review the questionnaire content. However, it would be much better if a full-scale pre-test could be conducted to test all aspects of the survey from planning through publication. This is discussed in more detail in paragraph 65.

25. The international recommendations published by the United Nations ^{5/} contain a brief discussion on the need for distributive-trade statistics. The following chapter attempts to describe the more important users of the information sought in the inquiries. It is hoped that consideration of the users as well as of the purposes for which the information is needed by them will assist the planners of an inquiry not only in selecting the items to be covered and deciding on their importance but also in preparing publicity and answering queries about the usefulness of the programme.

^{5/} International Recommendations on Statistics of the Distributive Trades and Services, Statistical Papers, Series M, No. 57 (United Nations publication, Sales No. E.75.XVII.9).

II. NEEDS AND USES OF DISTRIBUTIVE-TRADE STATISTICS

General remarks

26. Statistics on the distributive trades and services tend to be assembled much later in the course of economic development than statistics on agriculture, manufacturing and the extractive industries. This is so because the primary activities of man are initially involved in the provision of the basics, namely, food, shelter and clothing. It is only when society has been able to accumulate surpluses that trade becomes important. Specialization in the form of the provision of services is an even later development. Just as it was found useful to gather information concerning agriculture and manufacturing as an aid to understanding and improving these functions, so it becomes useful to gather data on distributive trades and services when these also become important economic activities.

Governmental needs and uses

27. There are many ways in which different departments of government can make use of data on the distributive trades and services. One of the most important is as a basis for assessing trends in the economy. For short-term trends, annual and more frequent estimates of this sector of the national economy play an important role. This is especially important in the fields of retail trade and personal services, since these activities reflect final consumption, the goal of all economic activity. In consumer-oriented economies the rise and fall of final consumption necessarily affects all other economic activity and thus becomes the single most important indicator to be watched by the governmental agencies responsible for economic policy. To the degree that production and consumption are not bridged through the medium of trade, this indicator becomes less important. In

addition to measuring final consumption, it is helpful to measure other aspects of the distributive trades and services, such as wages and salaries earned in the sector, changes in stocks and gross fixed capital formation.

28. Besides measuring trends in the economy, the data on distributive trades and services can be used in establishing policy (and in measuring the effects of policies already established) in many different areas of governmental activity, for example, in establishing the level of revenue that can be expected from the imposition of taxes such as sales taxes, value-added taxes and payroll taxes and in determining whether the expected level of income has been achieved. The data are also needed for the computation of national income and product statistics. Other programmes that may be helped include evaluating the need for modification of the transportation system, measuring the need for labour and its utilization, planning regional development programmes, drafting social welfare programmes such as social security, minimum wage and unemployment insurance, evaluating anti-trust programmes etc.

29. Despite the usefulness of data on distributive trades and services to the Government, it should always be kept in mind that supplying this information becomes an expense for the business firm which, in the aggregate, may be much higher than that for the Government. Requests for complicated data in this segment of the economy may be particularly burdensome, because the typical firm is small and may not keep adequate records. These facts should be weighed in determining the data content of the programme.

Business needs and uses

30. While the business community is the principal supplier of data on the distributive trades and services, it is also a very important user of such data.

It may also be a beneficiary of their proper use, because the decisions made by the Government as they affect business can be made more intelligently. Such decisions can be promoted or objected to by business representatives on a sound basis if the facts are available. While it is true that only larger firms can afford the staff to analyse the data, smaller firms can be effective in moulding policy through their trade associations in their role as representatives of small businesses.

31. One of the prime uses of the data by business is in the determination of the location of business outlets. Of course, data on the nature of population and income distribution need to be used in conjunction with business data. Industry analysts can get a good picture of an entire industry in terms of the size of units, the proximity to population centres, the ratio of payroll to sales and of inventory to sales, market penetration, industry trends over time, kind-of-business outlets for specific commodities and several other useful factors, depending on the kind of data collected and the time and energy invested in the analysis of the information assembled. The more complex information cannot, of course, be collected in the annual and more-frequent-than-annual surveys but only in the infrequent comprehensive censuses.

32. In addition to firms in a specific industry that make use of the data, firms in industries that perform services for business also find the data useful. For example, public utilities and transportation companies can do a much better job of planning if they have detailed knowledge of the structure of the many units involved in retail, wholesale and service trades. Similarly, all types of equipment manufacturers can study the market for their goods. Manufacturers of commodities that will be sold through wholesale or retail channels and wholesalers seeking knowledge of their retail outlets can make good use of the data. Financial institutions are another type of firm with a strong interest in the statistics.

Other needs and uses

33. A use of the data which should not be overlooked is that relating to the academic field. A primary application would be in schools of business administration as a basic tool in the training of personnel who will become part of management in moderately sized and large business organizations. Another application would be in the training of economists, market analysts and survey statisticians. Persons being trained as labour economists, financial consultants, or government economists will also find the data useful.

34. Labour organizations will find the data useful in their activities in the field of wage negotiations. They can relate data on their membership with information on the location and number of workers in the areas in which they are active to determine where their organizing efforts might be most effective. Furthermore, the knowledge they gain of the structure and functioning of the industry can be helpful in establishing reasonable demands and negotiating effective contracts.

35. A basic use to which the results of the infrequent comprehensive census can be put is in the design of the samples to be used for other surveys, especially those conducted annually or more frequently. The comprehensive census is the ideal framework from which the most efficient sample can be selected, since it contains such characteristics of the units to be sampled as kind, size and location of business and consists of a complete enumeration which includes the name and address of each unit. While private sampling organizations cannot have access to the detailed information available to the Government, it is still possible for them to make good use of the published data, both in the design of an efficient sample and in checking the results of their work against government totals.

36. Finally, the expansion of statistical programmes to include the dissemination of information concerning the distributive trades and services will add another facet to the general fund of knowledge and improve public awareness of the facts of business life which makes possible the informed influence of individuals on public policy. Those involved in moulding public opinion through various communications media--newspapers, magazines, television--will have access to the data which will permit better evaluation of public and private programmes.

Preparing a brief for approval of the programme

37. For the purposes of this manual, it is assumed at this point that some responsible persons in authority have tentatively decided that it would be useful to go ahead with a programme of statistics in the field of distributive trades and services. It is also assumed that there is already in existence a statistical organization which is engaged in other survey work. As an initial step in the process of receiving approval for the programme, it would be advisable to prepare a brief or justification for the proposed programme. Much of the work outlined below can be used in preparing such a brief. It should include some general idea as to cost, timing and personnel requirements. This need not be in great detail but should be adequate to give the agency head some indication of the order of magnitude so that he can make a decision as to whether to recommend approval or modification of the programme. A good job of presentation at this time will go a long way towards having the project approved.

Legislative requirements

38. Once the statistical agency has decided to sponsor the programme, it is usually necessary to pass a law (or to modify an existing statute) establishing authority for the programme. The law should be drafted to include reference to scope (which should be broad enough to cover all retail, wholesale and service trades), the frequency of the comprehensive censuses, the nature of the annual and more frequent surveys, the rules of confidentiality as they relate to the disclosure of information supplied by a business firm and the imposition of penalties for failure to respond or wilful falsification of the requested data.

This sets up the charter under which the various surveys can be planned for and conducted. Where wide powers to gather economic information are already vested in a government official, a decree or executive order should carry sufficient authority. The decree should contain the same provisions as those outlined above.

Determining organizational structure

39. Any decision on organizational structure will be conditioned by the existing statistical organization. Where there already exists a functionalized structure with separate subject-matter, planning staffs, a field organization, a publications division etc., it may merely be necessary to add a planning staff for distributive trades and services which would prepare the specifications for the various operating divisions which, in turn, would set up separate units for the handling of these surveys. Where, however, the processing functions have not been established as separate operations, it will be necessary to have an organization in which planning and processing are combined in a single group.

40. One function that is normally relegated to a separate staff is that of field work which in all probability will be performed for a number of different programmes. A field staff becomes especially important when the programme provides for data to be collected from an area sample in a series of current surveys. Where a country has not previously established a field organization, the initiation of such an operation becomes a major and costly task.

41. In weighing the merits of a system combining processing and planning functions in a single staff against those of a division of work based on the function to be performed, two important factors are the nature of the personnel available and the volume of work to be performed. The planning function requires personnel strongly grounded in subject-matter problems. The processing function also requires

specialized knowledge and skills. Ideally, the staff would be proficient in both types of activities, but this is often not the case. Furthermore, each activity will probably require the full time and attention of the personnel assigned to it.

42. It is very important to understand the effect that processing can have on the quality of the data that is to be produced. Too often, a staff whose chief function is to process data will use short-cuts and make procedural modifications which, while they may speed up the processing and reduce the cost, will so affect the data as to make them useless. On the other hand, combining the two functions may result in highly competent subject-matter personnel devising a very inefficient operation and the wasting of funds which results in the abandonment of the programme. It is extremely important that both functions be kept in mind at all times and that those engaged in planning always consider the way in which their plans will affect processing, and vice versa. A short course on subject-matter should be given to processing personnel which should be reciprocated by a short course in processing to be given to subject-matter personnel.

Assembling the planning staff

43. The functions to be performed by the planning staff consist of determining subject-matter content; consulting with data users, trade associations and respondents; designing and testing questionnaires; participating in planning data-collection and data-processing methods; preparing instructions to respondents; preparing specifications for processing reported data; resolving data-collection problems; planning the publications and writing texts; and reviewing tabulated data professionally. The staff to perform these functions must obviously be of a high calibre. The basic skills required are a knowledge of statistics and economics with some aptitude in mathematics.

44. A full-scale programme of both infrequent censuses and of frequent surveys would preferably require setting up two planning groups, one for each programme. This is based on the fact that a repetitive programme of frequent surveys, such as a monthly series, requires the full-time attention of the staff assigned to that work, and the assumption by that staff of the additional duties related to a comprehensive census would result in both programmes suffering.

45. The staff assigned to planning specific phases of the census should later have responsibility for carrying them out. This helps to ensure continuity of planning and implementation of the planned operations. It permits the training and experience acquired in the planning stage to be utilized to the fullest extent in the later census operations.

Organizing advisory groups

46. At an early stage, the planning staff should organize two major advisory groups of data users. These can be very helpful in utilizing the collective judgement and knowledge of government officials and the public in working out the objectives of the census. They can also be helpful in launching the census by helping to secure needed legislation, funds and public co-operation.

47. The first group that should be formed is a governmental advisory group. It should consist of experienced government officials who are users or potential users of the data. These officials should be in high-level positions, so that they can speak authoritatively for the agencies they represent. They should also be technicians who understand the nature of statistical programmes and their limitations. Possible candidates are representatives of the ministries of labour, social insurance, commerce and other agencies charged with regulating or promoting trade and services or economic planning and analysis.

48. The second major category of data users for which an advisory group should be formed consists of non-governmental users. These include representatives of market analysis groups (private and academic), representatives of communications media (newspapers, radio, television), statistical research organizations, economists and consumer organizations. Again, an effort should be made to select members who have some knowledge of survey problems and can make a contribution to the planning rather than people who cannot function in such a capacity. If it is found that the advisory committee is becoming unwieldy due to its size, it may be necessary to split it into two or more subgroups.

49. Two types of organizations have been deliberately excluded from the advisory groups, not because their advice is not needed, but because consultation with them is of a different nature. The first consists of trade associations which need to be consulted in greater depth and with specific reference to the needs and problems relating to a specific trade. It is, therefore, recommended that each trade association that represents an industry included in the survey be contacted separately to review the impact of the survey on that industry. The second excluded group consists of accountants. Representatives of this group need to be consulted with respect to data availability and the time needed for reporting. Again, this should be done as a separate activity, and accountants should not be included on advisory committees.

50. At the stage when the planning staff has arrived at a preliminary conclusion concerning the data content of the survey, it would be advisable to set up a system for keeping a formal record of all clearances conducted in arriving at the survey content. This record should include the date of contact, the person contacted, the method used (mail, personal interview, telephone), the comments of the reviewer and the response of the planning staff representative. Such a record will prove very useful for future reference.

Fundamental decisions

51. Once the survey has been approved and the planning staff has been assembled, there are several matters that must be decided, at least tentatively, at a very early stage by the planning staff. They are briefly examined in the following paragraphs.

Scope

52. While the general decision on the industries to be included in the survey will already have been made, it will still be necessary for the planning staff to review that decision with specific reference to each industry to be covered and to consider what modifications may need to be made. It may be established that with some slight additional cost the survey may be expanded to cover an industry in greater detail or to include an industry not previously contemplated. Or it might be determined that a particular industry should be omitted because it is impossible to enumerate it.

Coverage

53. A very important early decision to be made which has a great effect on the budget pertains to coverage. Especially important is the decision of whether to attempt complete coverage or to omit such categories as establishments operated in homes; those with no employees; those under a specified size or volume of business etc.

Availability of funds

54. The planning staff needs to know just how firm the commitment is to proceed with the survey and what level to plan for. The statistical agency should be able to make a fairly definite statement so that the objectives of the programme can be geared to a reasonable estimate of the funds that will be made available.

Method of assembling data

55. Once decisions are made on scope and coverage and it is reasonably certain that adequate funds will be made available, a firm decision should be made on the method to be used in assembling the data. Answers should be given to such questions as To what extent will data be obtained from other government agencies? To what extent will sampling be used? What portion of the universe will be covered by mail? To what extent will data be collected by field enumeration? How will physical location be determined for non-respondents? and How will addresses be obtained for mailing purposes? Representatives of both the planning and the processing staffs will have to work jointly on these questions.

Method of data processing

56. With the continuing advancement in the design of data-processing equipment, the number of options available for use in processing and tabulating data continues to expand. It is necessary to insert a word of warning at this point. Too often a decision is made to use sophisticated electronic equipment when it is really not only unnecessary but actually inadvisable. Even the most statistically advanced countries have had great difficulty using such equipment, and unless personnel are available who have mastered the problems of writing the programs, operating the machines and maintaining the equipment, it would be better to use the simpler systems. The first considerations should be the type of equipment available and the feasibility of using it for processing the results of the survey.

57. In any event, a decision must be made on the type of equipment to be used and the extent to which clerks and professional staff will be needed in editing reports and in preparing and reviewing tabulations. A related requirement is the evaluation of equipment and staff requirements for mailing, check-in and follow-up

operations and for the establishment and maintenance of a directory of business establishments.

Equipment and staff requirements

58. Once a general decision is made on the manner in which processing will be performed, it will be necessary to enumerate all the processing operations and when they will be performed. Estimated workloads and corresponding equipment and personnel requirements, by type, will need to be calculated. These should be co-ordinated with other agency requirements and new equipment ordered and personnel hired as needed. In estimating equipment requirements, allowance should be made for the fact that there is often a long lead time required for delivery. In addition, allowance should be made for time that will be lost due to the breakdown of equipment or for delays such as those due to the failure of a preceding operation to deliver a sufficient amount of work ready for processing

59. Where the statistical agency has established a centralized processing operation for systems-design, advantage can be taken of a staff specially trained to optimize processing operations. It is necessary, however, for the planning staff to review the procedures proposed by systems-design personnel, since it is not always obvious what effect a specific operation may have on the quality of the data that is produced.

Allocation of responsibilities

60. Because of the complex nature of these surveys and because implementation of the plan will affect a number of different organizational units of the statistical agency, some aspect of the programme can be easily overlooked. Listing all major activities and making sure that each unit involved becomes aware of its specific responsibilities with respect to that activity is, therefore, necessary. Each unit should review its operations to determine to what extent

its own work is dependent on the activities (such as writing specifications, preparing table plans, choosing the format of machine output, assessing the availability of test output etc.) of other units and make sure that each unit is fully aware of what is required from it and when.

Co-ordination with other surveys

61. While this manual is designed to be used in the organization and conduct of surveys of distributive trades and services, it should be remembered that such surveys are not conducted in isolation. They are frequently conducted jointly with surveys of manufacturing, mineral industries, construction etc. Since there are many common elements in all of these surveys and since a single multi-establishment enterprise may often operate establishments in a variety of industries, it becomes important to see to it that no contradictions exist in the reporting forms and instructions that are prepared by the different planning staffs engaged in survey work. Furthermore, in the preparation of mailing lists and in non-response follow-up operations, a jointly planned operation is not only more efficient but will aid in the resolution of problems of scope and in the definition of the reporting and statistical unit. It is, therefore, advisable to establish a "common elements committee" at an early stage in the planning which will include representatives of each division and will be responsible for determining the wording of the questions that will be common to all questionnaires. The same committee should be assigned the responsibility of preparing a joint mailing list and non-response follow-up plan.

62. While the wording for common questions need not be exactly the same on different forms, the effect that different wording may have on other surveys should not be overlooked. For example, in establishing the statistical unit for a manufacturing plant, there may need to be some reference to separate sales

locations which would be properly classified as sales branches and included in wholesale trade. The use of a common elements committee is of particular importance in the initiation of surveys of the distributive trades and services because such surveys are typically initiated subsequent to those conducted in the manufacturing industries. It will be found, in consequence, that there is a tendency for such problem areas as the identification of sales branches to be overlooked, resulting in an overstatement of manufacturing activity.

63. The charter of the common elements committee should cover such items as the wording of the questions on establishment identification, ownership, the legal form of organization, the period of operation and physical location. It should make sure that standard definitions are used for such terms as "employment," "payroll" and "capital expenditure". The committee should assume the function of planning and writing the specifications for a pre-canvass of multi-establishment enterprises aimed at determining the location, tentative industry classification and approximate size of each establishment operated by the enterprise, so that the proper reporting form may be sent. Another function is that of determining the content of the mailing label. The committee can also be used in liaison with other governmental agencies in matters that are common to all divisions, such as obtaining mailing lists or data.

64. It should be clear from the above that the members of this committee need to be highly competent technicians. There must also be on the committee representatives from the processing divisions to evaluate the feasibility and estimate the cost of the plans made by the committee. While the number of common elements involved in the conduct of frequent current surveys is much smaller than in comprehensive censuses, it is still very important to make sure that the questions and definitions are consistent in both. Problems

arise because the need for rapid reporting in current surveys makes it difficult for respondents to refine the data; in comprehensive censuses, the respondents have more time to develop accurate records. This can lead to inconsistencies between the two types of surveys. The planning staff for current surveys should, therefore, carefully review the work of the common elements committee.

Planning a pre-test

65. A very useful device for determining the feasibility and cost of a census is the comprehensive pre-test which would consist of taking an actual census of a representative portion of the country--perhaps an entire province--which contains both urban and rural areas. The ultimate purpose of such a pre-test should be the actual publication of data. This will necessitate careful planning and processing. Only in this way can one be sure of becoming aware of most of the problems likely to be encountered in surveying the entire country. Inevitably, there will be some problems specific to a particular industry or area which will not be met even in such a pre-test. While it may appear that spending time and money on a comprehensive pre-test is wasteful, the result in terms of the ability to refine procedures, train personnel and estimate a proper base for estimating costs should be a much more efficient census operation and greater assurance of its successful conclusion.

66. Another form of pre-test is useful in connexion with problems relating to a specific industry--for example, determining the correct wording for and feasibility of collecting data on occupancy rates of hotels or the number of admissions in theatres. In such cases, the preferable procedure might be to prepare a report form containing the specific items to be tested which would then be sent for completion to a representative group of firms in the industry. The response would then be analysed by the planning staff before the wording of the final question is completed. This type of pre-test can be used with equal effectiveness for both

comprehensive censuses and current surveys. In the case of frequent surveys, the early reporting periods should be considered as pre-tests. Plans should be made for a rapid intensive review of initial surveys so that any changes needed in procedures or on reporting forms may be made quickly.

Publication planning

67. The measure of success of a survey is its ability to produce the desired information within the available cost and time limits. Before undertaking a survey, the survey manager must know in detail what his data objectives are. The data users must have a chance to review the proposed output, and the processing staff must know what goals it is trying to achieve. The best way to attain these objectives is for the planning staff to prepare a set of proposed tabulations for review and comment by the data users and processing staff.

68. A tentative publication plan, if prepared thoroughly and thoughtfully, will force the planning staff to ask some hard but important questions such as Is the questionnaire becoming too long and complicated because of the special interests of certain data users? Are questions being worded so that the answers are not mutually exclusive? Can all of the questions be tabulated? Will a small number of respondents for a particular cell of data result in too many suppressions because of disclosure problems? Will the proposed tabulation result in an impossibly large number of pages of data to be published? Can data from two or more tables be combined in a single table, so that money is saved? Will the amount of machine time needed to produce a proposed table be excessive? Is the total publication programme so elaborate as to be too costly or time consuming? and What will be the effect of sampling variance on the ability to publish information? These and other questions that require answers will come to light during the review of the publication plan.

69. The processing staff must be kept advised of all proposed tabulations so that they may evaluate the proposals in terms of the cost and time it will take to prepare them. This should be done at an early stage so that major problems may be discovered and alternative solutions explored. Once a final set of tabulations is agreed upon, work can proceed simultaneously on a variety of fronts. These include the designation of the report form content, the assignment of codes to appropriate data items, the keying of the information to be punched, the preparation of clerical review procedures, the design of machine processing procedures for data assembly and analysis and the preparation of tabulation plans. At this stage it will be possible to refine the original cost, personnel and time estimates and to make any adjustments in the programme that may be necessitated by this evaluation.

Differences between planning for census programmes
and planning for survey programmes

70. It is possible to use the same planning staff for both the comprehensive census and the current surveys, or to have two separate staffs. Both approaches will have some defects and some advantages. The major defects of having separate staffs consist of the inconsistencies that develop between the two programmes because different people are involved, with resulting communication problems, and the problem of trying to maintain a trained census staff throughout the census cycle, with resulting budgetary difficulties. On the other hand, having a single staff usually results in the census being inadequately staffed, since the need to publish many periodic reports means that the major effort of the staff is inevitably concentrated on that activity.

71. In planning for frequent surveys, it is possible to establish a regular cycle of operations with a known and predictable work load. This means that the size of the staff and the cost of the operation can be established and that the level of

trained personnel available can be fairly well maintained. This does not apply, however, when it becomes necessary to revise the sample, which is usually after the completion of a comprehensive census from which updated information on classification, size and panel membership must be extracted from census files. It is possible to use some of the census planning staff which should become available at this time to help in this operation. However, there will be very large additional costs incurred for field operations and processing.

72. The size of the staff required to process the infrequent census is subject to wide variation over the census cycle period, particularly in the case of processing personnel and equipment. This may lead to inefficiencies in the use of equipment and to problems in obtaining the clerical staff for the comparative short period required; therefore, it becomes especially important to schedule operations closely. When a combination of mail and field canvass is used in a comprehensive census, there are the additional problems of establishing field offices and hiring and training enumerators for a comparatively short period of time. This operation is even more difficult and expensive than establishing a temporary clerical processing office. Unlike the repetitive surveys in which enumerators, once trained, can be used over and over again, a comprehensive census requires that the enumerator be able to do his job well the first time, since there is very little time for him to benefit from correction of his mistakes.

IV. MANAGEMENT AND BUDGET

Use of manuals

73. One of the most useful ways of establishing the rules under which a survey is to be taken and the procedures to be used is to set up a series of formal manuals to be distributed to all persons responsible for planning and processing. The manuals should be designed to be incorporated into loose-leaf binders, and the numbering scheme should be flexible in order to permit expansion of each section. Provision should be made of a means to substitute pages when specifications or procedures are changed after the initial section has been incorporated into the manual. Each section should be dated as issued, and changes should also be dated. Those sections that have been superseded by later versions should be removed from the main body of the manual and retained in a separate section for reference, if necessary.

74. Basically, the planning staff will be responsible for writing specifications telling the processing staff what needs to be done, and the processing staff will write procedures on how to accomplish these goals. Frequently, however, in order to assure attainment of the desired goals, the planning staff will have to write the specifications in such detail that they almost constitute a procedure. The determination of the most appropriate techniques for achieving a specified objective should rest primarily with the processing staff, but both the final specifications and the final procedures should be established after a review of draft versions by members of both staffs.

75. The manual set up by the planning staff should consist of separate chapters covering such subjects as:

Mailing list assembly
Mail out, check-in and follow-up
Administrative records
Field operations
Sampling
Coding schemes
Clerical processing
Machine processing
Publication plans.

Use of flow charts

76. Another device which can be used in conjunction with or as part of the manual is the presentation of specifications and procedures in the form of flow charts. A survey can be visualized as a series of interdependent actions leading to the publication of data. Each action stems from and flows into another action. The relationship between these actions can be presented either diagrammatically or verbally. Both methods can be useful. The diagrammatic presentation in the form of a flow chart has the advantage of being more concise and makes it easier for many people to follow readily the flow of work; it also makes it easier to discover illogical steps or the omission of necessary operations. It is possible to present in a one-page flow chart material that in written form might take many pages.

Time schedules

77. One of the first steps in setting up a realistic plan of work is to establish a time schedule for the completion of key activities. Since the objective of the survey is to produce data, the first date to be set is that for the production of publications. Once that date is established, it will be necessary to work backwards through all the operations that are needed to achieve the goal. At this stage, only key dates should be determined. After such key dates are determined, it will be necessary to see what needs to be undertaken to meet this schedule. Some of the key operations, in addition to publishing the data,

for which dates must be established are (in reverse order) planning the tabulation, machine-editing the data, processing the forms, mailing the report forms, establishing the mailing list, performing sampling procedures, conducting field operations, assuring the availability of report forms, determining the content of forms and establishing scope and coverage.

78. In establishing completion dates for an operation, it is also necessary to determine when the operation needs to be initiated in order to complete it on schedule and what other operations need to be completed before it can be started. It can then be ascertained whether sufficient lead time has been provided to avoid delaying subsequent operations. Establishing dates for the completion of key operations, assuring that sufficient lead time has been provided to complete such operations on schedule and checking that preceding operations will be completed on time can go a long way towards making a survey successful. All the other operations that revolve around these goals and the planning of staff and equipment use can then be undertaken. If there is any slippage in the achievement of these key dates, subsequent operations will need to be re-evaluated and dates changed or resources shifted to keep the survey on schedule.

79. A picture of the major operations can be charted by listing them and indicating beginning and completion dates in a series of columns representing calendar quarters. This type of presentation will make obvious any overlapping operations and those operations that need to be rescheduled because they will not be completed when needed. It also becomes the basis for establishing work loads by time period and, therefore, of indicating personnel and equipment requirements by time period.

Scheduling staff and equipment use

80. On the basis of the chart described above, each organizational unit that has responsibility for a specific function should prepare an estimate, by quarter, of how much and what type of personnel and equipment will be needed to achieve the indicated objective. In the preparation of this estimate the unit should divide the major operations for which it is responsible into suboperations for which time objectives should also be indicated. This may reveal potential bottle-necks or excessively costly operations. When all the estimates are aggregated, it will be possible to get the over-all picture of cost and personnel requirements by time period. After making any adjustments deemed desirable, a firm schedule can be established, and all units will be aware of their responsibilities. Each unit should, however, be given the opportunity to present any objections or reservations it may have concerning the goals set for it, and these should be thoroughly discussed before final decisions are made.

Progress reporting system

81. The proper management of a survey is dependent on knowledge of whether work is proceeding within time and cost estimates. Once objectives have been set, a system of progress reporting must be installed.

Unfortunately, not all operations lend themselves to ready measurement of progress. Functions such as key punching, clerical review of reports and answering correspondence are comparatively easy, because the activity is fairly uniform for each of the units processed and the work load, production and backlog can be stated in terms of such units. Each person engaged in the process can be judged on the basis of his production rate and evaluated by comparison with the production rates of other clerks. In addition, over-all production rates provide the basis for determining whether the activity is proceeding on schedule. If necessary

procedures can easily be changed to increase productivity, or inefficient clerks can be dismissed or transferred. On the other hand, such activities as writing specifications for clerical or machine processing do not lend themselves so well to progress measurement. However, it is still necessary to set time objectives and to follow progress closely on such activities. It is especially important to monitor them closely because it is usually not possible to compensate for falling behind schedule by assigning additional staff, since the type of personnel involved is highly specialized.

82. The most difficult function for which to evaluate progress is that of writing and proving computer programs. This is because only after a program is proved to be correct by being run with real data can one be sure that the work is completed. Any difficulty that appears even in the final stages may take a great deal of time to correct.

83. Progress should be measured in two ways--units produced (as related to final goals) and cost engendered. While units produced can be measured and presented on a time basis, information on costs is produced with greater difficulty and usually lags too far behind to be of practical value. A usable approximation of cost, however, can be made by collecting data on man-days of work performed, by function, and assigning a calculated cost per man-day, which should be estimated separately for each function. Similarly, costs can be assigned to such functions as computer operation by accumulating daily information on the number of computer hours used and multiplying them by the estimated cost per hour. Both units completed and accumulated cost should be analysed frequently in terms of scheduled completion dates and funds allotted for the function, and corrective action taken early enough to be effective.

84. At least two levels of progress reporting are required. For over-all management, reporting can be confined to major operations and cost elements. For lower-level management these major operations should be divided into significant suboperations relating both to cost elements and to those operations that could become bottle-necks, even though they may not involve great expenditures. It must be kept in mind that costs incurred because an operation takes longer than originally scheduled affect not only the operation in question but frequently result in idling personnel and equipment assigned to later operations, with mounting attendant costs.

Establishing a quality-control system

85. In addition to keeping operations on schedule, it is very important to maintain acceptable quality. This is especially important in a comprehensive census, since there is no opportunity to correct an error in subsequent surveys as there is in a survey that is repeated monthly or quarterly. To monitor operations, therefore, a quality-control unit should be organized to design a system of samples to be selected of various operations so as to establish whether acceptable levels of quality are being maintained. In samples relating to key punching and clerical operations, the system can be designed to determine also whether individual clerks are meeting acceptable levels of quantity. Quality-control systems can also be designed to check on such things as computer-prepared mailing lists, field operations, the proper response to correspondence, coding etc. The staff of the quality-control unit should be independent--that is, it should not be an integral part of the unit it is supposed to be monitoring. Before quality-control procedures are activated they should be reviewed by processing and planning staffs.

Maximizing the use of resources

86. There are several devices that can be useful in stretching available resources. One is a judicious use of overtime. Often a manager, in trying to achieve a time objective, has to weigh the alternative costs of training a larger staff against using a smaller staff and investing in overtime work, even though at higher rates. The latter alternative may turn out to be less costly. Overtime, however, should never be used so heavily as to result in inefficiency because the employee becomes too fatigued.

87. Advantage can be taken of the fact that many operations in a census are sequential to use the same personnel in different stages of the operation. The training and experience gained in one operation is transferred to later operations and results in better and more efficient work. To achieve this requires very careful planning and scheduling as well as careful adherence to time schedules.

88. A statistical agency that is engaged in processing a number of different surveys will find that the use of mechanical and electronic equipment fluctuates greatly. Good scheduling can take advantage of periods of low usage and result in a general lowering of costs. This is particularly true with respect to computer usage.

89. One of the major problems that has to be faced affecting cost is controlling the flow of paper. Directly related to it is the question of how to make optimum use of clerks. A balance must be struck between the effect of simplifying each clerk's function to the point where he can perform it with extreme rapidity and the problems arising from the resultant increase in the flow of paper from one clerk to another. In general, the operations to be performed by a clerk should include as much as he can readily learn to handle effectively in a limited training period. Where numerous and complicated controls must be established because paper is being excessively transferred from one operating unit to another, there is loss

of efficiency, frequently aggravated by difficulty, or even loss, of control.

90. Costs can be controlled to some degree, especially in a comprehensive census, by the judicious use of sampling for those data items for which publication is required only at national or broad regional levels. This will complicate the survey to some extent because of the need to design samples and the attendant problems of processing them, but a net gain can be achieved in operating costs.

91. Where mass clerical operations are being conducted, it is important to establish a set of standards against which the production of individual clerks can be measured. These should be reviewed periodically and posted, perhaps weekly, so that each clerk can be aware of his relative position. Related to production rating must be quality rating, since production achieved by loss of quality is self-defeating. A clerk whose production falls consistently below acceptable levels should either be dismissed or transferred to another operation if it is felt he could do an acceptable job elsewhere. This serves the double purpose of eliminating unproductive workers and of serving as an incentive for higher achievement.

92. Regardless of how well the survey is planned, the fact that there are so many activities and that responsibility for different phases of the survey may be scattered among different divisions of the statistical agency frequently results in the creation of bottle-necks, the need for resolution of conflicting opinions about a course of action, the necessity to shift personnel, the changing of priorities, the adjusting of budgets etc. It is for situations such as these that a person with sufficient authority should be appointed to co-ordinate the survey operation and to anticipate and resolve problems and conflicts. While much of the information he needs to have to function effectively will be derived from various operational progress and budget reports, he should also be free to

inspect any operation personally and to probe to find solutions for problem situations before they deteriorate too far. He should be free to call meetings of personnel representing different divisions, to adjust priorities and to conduct negotiations with other divisions aimed at keeping the programme on schedule and within the budget.

Differences between censuses and current surveys

93. The fact that current surveys are repeated at regular, short intervals, whereas a complete census of distributive trades and services should probably be planned to be conducted every 5 or 10 years, has a marked effect on budget and management. Costs for the repetitive surveys tend to be stable from month to month, and operations can be refined over time. Schedules that can be pinpointed to the day of the week and even to the hour of the day can be established, when necessary. This is not true for censuses, where the great bulk of the work and cost falls in the year following the reference period of the census and the operations and budgets for each of the other years vary greatly. In current surveys, the emphasis on early and regular production of key measures means adherence to very tight schedules which can be planned for by processing personnel, even to the extent of regular, planned periods of overtime work. In a census, everything must be geared to getting the work done properly the first time; otherwise, it must be redone until errors have been corrected before the next operation is undertaken. When an error is not detected until several additional operations have been completed, all intermediate operations may have to be redone. This can be very costly.

94. The changing cost of labour and other needs due to wage and price increases over the normal 5 or 10 year cycle of a census can rapidly invalidate budget estimates. Fluctuation in the availability of personnel over such a period may also cause difficulties. The loss of trained personnel from one census period

to the next is another great disadvantage which results from changes in manpower requirements not only from month to month but also from census to census. This is particularly bad when there is a 10 year period between censuses.

V. SCOPE, COVERAGE AND CLASSIFICATION

Introduction

95. Two United Nations publications should be considered integral parts of this manual; they are International Standard Industrial Classification of All Economic Activities (Sales No. E.68.VIII.8) and International Recommendations on Statistics of the Distributive Trades and Services (Sales No. E.75.XVIII.9).

96. The International Standard Industrial Classification (ISIC) contains a comprehensive discussion of the principles used in establishing the classification system and how to apply those principles. The discussion includes sections on the purpose and nature of the system, the criteria used in constructing it, the identification of statistical units, the classification of the statistical unit, the structure of the system and the adaptation of the international system to national requirements. The international recommendations include specific recommendations on the scope and coverage of the proposed surveys; a discussion of the statistical unit; the characteristics that should be determined for each statistical unit; the items of data (and their definitions) that should be collected; and a set of tables which contain recommendations on what data items should be included in infrequent inquiries, in annual inquiries and in more-frequent-than-annual inquiries, with separate recommendations for countries with already developed statistics and countries in the process of developing statistics in the field of distributive trades and services.

97. The following presentation is not intended to substitute for the above-mentioned publications but should serve to point out major areas of difficulty

and to suggest methods of handling such difficulties. It is recognized that the economies of different countries will vary considerably and that there are national considerations that will condition the nature of the specific actions that will be taken. It is, however, hoped that attempts will be made to conform as much as possible to the recommendations, so that international objectives may be achieved. Whenever possible, the following discussion will attempt to highlight differences that may be relevant when annual and more-frequent-than-annual surveys are considered in contrast to comprehensive censuses.

Scope

Wholesale trade (ISIC 61)

98. The international recommendations suggest the inclusion of all types of wholesalers as part of the scope of a survey of distributive trades and services. Where a comprehensive field enumeration designed to identify every business establishment is undertaken, this objective is attainable, although some difficulty will be encountered in locating those businesses that are operated directly from homes, unless their addresses are available from business licence lists or other administrative records. When, however, the survey design calls for the use of a combination of a mailing list of larger firms and an area sample to represent smaller firms (such as is common in more-frequent-than-annual), difficulties develop which call for a modification of the scope of that portion of wholesale trade that represents non-merchant wholesalers.

99. There are several problems. First, the number of non-merchant wholesalers is comparatively small, and they are not distributed physically in such a pattern as to make them susceptible to effective area sampling. Secondly, such non-merchant wholesalers as sales agents and brokers tend to have rather small businesses and frequently operate from their homes. This makes it difficult not

only to locate them in a field enumeration but also to draw up a reliable mailing list. Manufacturers' sales branches and offices present another type of problem. Unless a survey of manufacturing is being conducted simultaneously with one of wholesale trade, the identification and proper enumeration of such branches and offices are very difficult.

100. Therefore, for annual and more-frequent-than-annual surveys of wholesale trade, it is probably better to restrict the scope to merchant wholesalers, defined in ISIC as "wholesalers who take title to the goods they sell, such as wholesale merchants or jobbers, industrial distributors, exporters, importers, terminal elevators and co-operative buying associations". If, however, up-to-date lists of all business firms along with their kind-of-business classification and some indication of size can be maintained, they can be used for either a complete mailing or the selection of a representative sample. However, it is rare for this set of circumstances to prevail.

Retail trade (ISIC 62)

101. This aspect of distribution tends to follow fairly closely the pattern of population distribution and thus lends itself readily to the use of an efficient area sample design. The distribution problem that exists with wholesale trade does not arise, and it is, therefore, possible to include all types of retail trade in those annual and more-frequent-than-annual surveys that incorporate an area-sample feature. Because of the large percentage of retailers that go into and out of business within a short period of time, it is very difficult to develop a mailing panel that, by itself, can be used for annual and more-frequent-than-annual surveys. Such a panel tends to become representative of only the larger, more stable firms. It is, therefore, better to develop an area sample that can be used to represent all smaller businesses as a supplement to the list of large establishments usually derived from a comprehensive census.

Service trades

102. The scope of the service trades does not lend itself to ready definition. Furthermore, the international recommendations for this group of trades is highly selective. It is designed to be confined to "services rendered to households and enterprises of the types normally provided by establishments which can be described as stores, hotels, restaurants, theatres, shops and offices". There are, however, a number of exceptions, and the proposed scope and attendant qualifications should be studied carefully. An inspection of the nature of these activities will indicate that some of them will present great problems if they are included in annual or more-frequent-than-annual surveys, especially if up-to-date mailing lists cannot be developed and kept current. In several instances (seasonal hotels, motion picture production, radio and television broadcasting), the use of area samples would be unwise.

Coverage

General

103. It can be assumed that, in practice, complete coverage of the retail, wholesale and service trades is virtually impossible. This is so because many of the establishments doing business in those kinds of activities operate in a manner that makes it very difficult to find them, even during the course of field canvassing, and many will never appear in any tax or administrative record system. It would be desirable to make certain distinctions in coverage between the infrequent, annual and more-frequent-than-annual inquiries. While the infrequent inquiries should, in principle, cover all retail, wholesale and related service establishments - irrespective of size, whether or not recognizable from the outside, whether or not fixed in location, it is not practical to attempt this wide coverage in the annual and more-frequent-than-annual inquiries. Those inquiries, to be feasible, should be restricted

to units with fixed locations, recognizable from the outside, at least in the case of the retail and service trades. Recognizability is probably too restrictive in the case of wholesale trades, and it may be feasible to build up suitable directories without field canvassing each time. The small retail and service units will, however, require field canvassing of the sample areas each time.

104. The major types of problems that affect coverage are: (a) non-recognizable establishments, particularly those that operate in homes; (b) itinerant establishments, those that move about, with no fixed place of business; and (c) temporary stands which may be open only certain days of the week or seasons of the year. The survey manager must decide whether, in his particular case, the effort needed to identify and enumerate such establishments is warranted or whether he can legitimately use such devices as arbitrary size cut-offs or area samples to reduce his problem.

Size cut-offs

105. In most cases, the three types of establishments mentioned will undoubtedly be small in size. As a result, it may be assumed that a size can be determined below which the great majority of those establishments will fall. If a rule is made that all establishments below this cut-off will be omitted from the survey and that no attempt will be made to find or enumerate non-recognizable, itinerant or temporary establishments, then the universe can be defined as all those over the cut-off size. Of course, where a substantial portion of a country's activity is performed by these smaller establishments, the effect on the statistics can be quite marked. A country may then decide to cover all recognizable places of business, regardless of size. This should be clearly indicated in the report, as should any size cut-off that is used. In both instances, for the guidance of the data user, some effort should be made to evaluate the amount of loss in coverage.

Area sampling

106. In the kinds of activities that tend to be distributed in proportion to

population, such as retail trade, eating and drinking places and many of the service trades, it is possible to design an efficient area sample that will reduce the number of small establishments that will need to be contacted. This may make it possible to devote more effort to identifying and enumerating the establishments that are more difficult to find, but the use of area sampling will greatly reduce the amount of kind-of-activity and area data that can be published at reasonable levels of sampling variance.

Special problems in field enumeration

107. The specific difficulties so far referred to--non-recognizable places, itinerant vendors and temporary places of business--are especially significant in a field enumeration. The following additional difficulties may affect coverage when this type of procedure is used:

(a) Places closed at the time of enumeration. When a place of business closed at the time of enumeration, the enumerator must establish whether the establishment is closed temporarily, is out of business or is open only during certain hours or certain seasons of the year. Getting reports for this type of establishment may be quite difficult. In some cases, such as in the enumeration in resort areas or of seasonal hotels, this problem can be especially important;

(b) Firms operated in out-of-scope establishments. There may be a tendency to overlook in-scope activities that are located within establishments that are not in the scope of the distributive trades and services, such as museums, railway stations, educational institutions, apartment buildings, professional buildings etc. The in-scope activities found in such locations should be listed and enumerated only if they are open to the general public and are privately operated;

(c) Leased departments and concessions. It is easy to overlook firms

that operate as leased departments or concessions, such as a watch repair firm in a jewellery store, a beauty shop in a department store, a meat counter in a grocery store. This could significantly affect coverage in certain kinds of business and may even lead to a general understatement of total retail sales;

(d) Shared office space. In order to reduce operating costs, firms sometimes rent "desk space" in another firm. They may or may not be identifiable, but firms such as agents and brokers can easily be overlooked by the enumerator. Such firms may account for a sizable portion of the business done by export agents, grain dealers, import agents, food brokers etc.;

(e) Mail-order houses and vending machine operators. These types of establishments have no need to attract customers to their places of business and, therefore, the enumerator may tend to overlook them. The mail-order house makes contact with his customers through direct advertising or through magazines and newspapers; the vending machine operator deals directly with the establishments in which he places his machines. Special efforts need to be made to ensure coverage of these types of firms.

Special problems in mail enumeration

108. In mail enumeration, the use of a directory, particularly one developed from administrative records, avoids some of the problems encountered in field enumeration but has some drawbacks of its own:

(a) In-scope units of out-of-scope firms. Some of the larger firms will operate many units, some of which will be in-scope and some out-of-scope of the distributive trades and services. For example, in administrative records the firm may be classified as a railroad and

thus thought to be out-of-scope. However, such a firm may also own and operate hotels and restaurants. With the spread of conglomerate and multinational companies, it is important to identify each unit of a company which is within the scope of the survey. Prior to the actual enumeration, therefore, a pre-canvass of a large firms should be undertaken to determine whether any of their establishments falls within the scope;

(b) Out-of-scope units of in-scope firms. In a similar manner, it is important to exclude from the survey those units of in-scope firms which are not in the scope of the census, such as a manufacturing plant owned by a retail chain organization. Again, the way to accomplish this is through the pre-canvass of large organizations;

(c) Limitations of administrative records. Because the administrative jurisdiction of an industry may be limited to firms of a certain size or with special characteristics, the coverage may not be as complete as that needed for survey purposes. In addition, the kind-of-activity classification may not conform exactly to ISIC criteria. The effect on the resultant statistics should be researched, and if it is decided to use administrative records in a mail survey, the final publication should include a discussion of any significant undercoverage.

Non-profit and government operations

109. In many countries a great many of the transactions taking place in a particular kind of activity are conducted by non-profit or governmental organizations. Those organizations operate in the market-place in a fashion similar to that of private business organizations, for example, the operation of farm producer co-operatives for the marketing of milk or grain and liquor retailing by government

agencies. To get a complete picture of the distributive trades, it is necessary to cover such activities along with those operated by the private sector.

Classification

110. As mentioned above, ISIC contains a very complete and detailed discussion of the principles of classification. In addition, it includes a discussion of the various types of statistical and reporting units for which data can be appropriately gathered. It would be pointless to repeat this discussion. It should be mentioned, however, that the principles of classification and of identifying reporting and statistical units must be mastered in terms of the situation prevailing in the country in which the survey is to be made, and the questions that will elicit the needed information must be designed with extreme care.

VI. QUESTIONNAIRES AND OTHER DATA-COLLECTION FORMS

111. As mentioned above, the international recommendations on statistics of the distributive trades and services contains a detailed presentation of data and definitions of the items of data that should be included. The choice of items and of the priorities are based on a consideration of the relative usefulness of, or the need for, the statistics and the difficulties in gathering the items. It is recognized that needs and circumstances will vary from country to country and that some modifications will be required in specific cases. The discussion below is confined to some of the problems that may be encountered in implementing the recommended definitions and in designing the questionnaire and other forms that would be helpful in conducting the surveys.

Information needed for control, correspondence and imputation

112. Every data-collection form should ask for the name and mailing address of the owning firm in addition to the physical location of the establishment, so that mail can be sent to the firm concerning any questions that require correspondence after the completed questionnaire has been received and edited. (The mailing address is frequently different from the physical location.) When a directory is used for preparing the address record, it is extremely helpful to incorporate codes indicating kind of activity and size in the address label. This information serves two purposes: first, it permits imputation for late respondents and delinquents, and secondly, it serves to indicate how strong a follow-up effort needs to be made for delinquents and incomplete reports. (A small firm with no employees obviously needs less follow-up effort than a firm with 100 employees.)

113. In addition, the mailing list should contain a unique identification

number for each establishment which can be used as a control number for the check-in of reports, the follow-up of delinquents, the filing of report forms and general control purposes.

Questions for publication and editing

114. It is obvious that the report form must include questions that will elicit the data that will be published. What is not so obvious is that a number of additional questions must be included which are designed to make sure that the data are properly reported and processed. The following are some examples:

(a) If sales and excise taxes are collected by a retailer and it is desired to have them included as part of the total retail sales, it is not always possible to get the correct response by merely asking the respondent to include such taxes. The safer method would be to tell him to include them, then ask him whether they have been included and then, if the answer is "no", ask how much was sent to taxing agencies. The editor then has all the information he needs to arrive at the correct answer;

(b) Frequently, more than one firm may be operating under a single name at a single location; for example, a department store with concessions for the operation of a perfume or shoe department. It will be necessary to design a question that will elicit this type of information and obtain sufficient information to permit proper handling of it;

(c) The inclusion of "self-editing" questions can be very useful. For example, the respondent can be asked to check the relation between his payroll and receipts to see whether they are consistent. Or he can be asked to check consistency in the totals that should occur in more than one place on the report form - once, perhaps, as a total sales figure

and a second time as a breakdown of sales by merchandise line. Such checks should be kept simple so that they can be done rapidly;

(d) For annual and infrequent surveys, the international recommendations indicate a preference for calendar-year reporting. However, if a fiscal-year period is normal for most establishments, then data may be compiled uniformly on that basis. While the questionnaire may indicate which method is preferred, it will be necessary to ask the respondent which method was actually used and, for fiscal-year reporters, information on the period covered by the fiscal year will also be needed, since not all firms will use the same fiscal period. The acceptance of information for a period other than that requested should be reviewed in light of its effects on the results and, if significant, efforts should be made to obtain key information, such as employment or sales for the required period, which can then be used to pro-rate other significant items of data;

(e) A special problem occurs in more-frequent-than-annual surveys when the firm operates only part of the year, as is the case for a seasonal hotel or an establishment located in a seasonal resort. It is frequently not possible to distinguish such an establishment from one that has gone out of business or from a vacant location unless the enumerator obtains answers to a series of questions. The omission of seasonal or temporarily closed establishments could result in a markedly downward bias.

Differentiating non-reporting entries from zero entries

115. A common problem in the processing of report forms occurs when a respondent fails to make an entry for an item for which the response could possibly be "zero"

or "none". The editor is then faced with the dilemma of determining whether this is a non-response or whether the respondent made no entry because he thought it was not necessary. While it is helpful to include an instruction that tells the respondents to make an entry of "zero", "none", or "-" in every case where appropriate, many will not follow the instruction. It is possible to reduce the impact of this problem by asking a question with a simple "yes-no" answer and using the response to establish whether to make an entry of "zero" or to process the question as one for which data were omitted.

Designing the kind-of-activity question

116. The presentation of data by kind of activity is one of the basic goals of surveys of the distributive trades and services. To arrive at a proper classification in accordance with the standards established in ISIC requires the application of certain objective criteria which, most frequently, requires obtaining answers to a standard set of questions. There are some kinds of activity, such as barber shops, radio broadcasting stations or motion picture producers, where self-designation may be acceptable, but in most cases it is necessary to probe deeper to get correct classification. For example, "clothing business" is insufficient for classification, since it may mean manufacturing, wholesaling, retailing or repairing clothing. Furthermore, the respondent may think of himself as a retailer when, under ISIC, he would be considered a wholesaler. The kind-of-activity question, therefore, must include questions that will permit classifying the establishment first as to whether it is merchandising, manufacturing or engaged in one of the service trades. If it is in merchandising, is it primarily engaged in retailing or wholesaling? If it is providing services, such as repair services, is that its primary activity? If it is engaged in merchandising and the publication programme calls for tabulating according to a detailed breakdown of activity, additional questions will need to be asked on the distribution of sales, by the kind of merchandise sold.

117. If a source for a preliminary classification is available, it is possible to use a specially designed kind-of-activity question tailored to the general kind of activity of the establishment. Thus, it is possible to have a special question for firms engaged in the food business, in the clothing business etc. This permits the use of pertinent questions and terminology and makes the questionnaire more meaningful to the respondent. It also permits classification by a more detailed breakdown of activity and the asking of specific questions pertinent to the general kind of activity, such as, for example, the sale of frozen food in food stores.

Detecting consolidated reporting

118. Because of the difficulty of identifying all multi-establishment enterprises prior to conducting mail surveys, some method for discovering such firms at the time of enumeration is needed. For this purpose, the report form should include a request for a list of all the business locations of the firm and basic information for each, such as kind of activity (self-designated), volume of sales, annual payroll and number of employees. Since, presumably, the larger enterprises will have been pre-identified and have had separate data requested for each location, the number and size of multi-establishment firms added at the time of enumeration should be small.

Keying the report form

119. A key code should be assigned to each item of data or other type of information that will be tabulated or used for sorting or that is necessary for the identification of the establishment. The type of code to be assigned can vary, depending on the kind of equipment to be used, such as book-keeping machine, punch-card equipment or computer. The proper assignment of codes can make a material difference in the cost and quality of work and should be done with knowledge of how the specific item of data will be processed. For example, where punch-card equipment is to be

used, care should be taken to ensure to the extent possible that those items of data that will appear in the same table are included on the same punch-card.

Importance of pre-coding

120. A great deal of time and money can be saved by examining each item on the report form to see whether it is possible to design questions in such a manner that pre-assigned codes can be printed directly on the form instead of having a clerk assign such codes. This has the double advantage of making it unnecessary for the clerk to take the time to determine and enter the code and of making it easier for the puncher to find and read the pre-printed code rather than depending on the sometimes illegible script of the coder.

Treatment of instructions

121. When a questionnaire is completed by a field enumerator, instructions are usually contained in a manual, and time can be taken to train and test him to ensure that the instructions are understood. If, however, the enumerator must leave the questionnaire with the respondent or if the survey is conducted by mail, then the instructions on, or accompanying, the form must be depended upon to elicit the required information. When a combination of field enumeration and mail is used, bias can be introduced if the instructions used by the enumerator are different from those on the mailed form. When different groups of people are responsible for the census and for the annual and more-frequent-than-annual programmes, special care must be taken to co-ordinate instructions. Lack of such co-ordination can easily lead to non-comparability of data in the two programmes.

122. The instructions for respondents should cover, precisely but completely, everything necessary to explain what is required and to avoid misunderstanding. Instructions that are long and diffuse are less likely to be read and understood than ones that are brief and to the point. If long instructions are required, however,

a common practice is to include only the most essential points on the questionnaire itself and the remainder on a supplementary instruction sheet. Pre-survey interviews with potential respondents often provide valuable indications of the points that should be emphasized in the instruction.

123. The following is a check-list of points normally covered in instructions on questionnaires or on supplementary instruction sheets:

1. Legal authority for the survey
2. Confidentiality of returns
3. Types of establishments required to report
4. Acceptability of estimates where accounting records are not available
5. Date the completed report is due
6. How to return the completed report
7. Time period covered by the survey and rules concerning treatment of fiscal-year data
8. Definitions of the various items of information requested
9. Instructions on the compilation of special measures such as percentage of occupancy for hotels or attendance at theatres
10. Special instructions for particular industries where problems affecting many respondents are expected, such as the treatment of leased departments in department stores.

Problems arising from the diversity of activities

124. The great diversity of activities included within the scope of the distributive trades and services creates difficulties, since arriving at a proper classification requires the answering of a number of questions, some of which may be relevant to only a small portion of the universe. In addition, there are special data questions that apply to specific kinds of activity. If only a single form were to be used, the resultant length and complexity would confuse and irritate respondents and probably greatly reduce response rates. The approaches to resolving this

problem that are used in surveys involving field interviews should be different from those used in surveys handled by mail. It is assumed, however, that a number of different questionnaires have been designed, each of which would be used by establishments which would be asked the same questions, for example, food stores, apparel stores, amusement places, petroleum wholesalers etc.

125. In surveys that involve a field interview (whether in connexion with complete censuses or in sample areas), the initial interview should include a series of questions designed to determine whether the establishment was a manufacturing, wholesaling, retailing or service one and establish within each such major segment which report form should be completed. The enumerator can then either leave the form to be completed and returned to the statistical office, or the listing sheet can be used as a basis for having the statistical office send the correct form to the establishment.

126. In mail surveys, it will be necessary to have enough information in the mailing records to mail the proper questionnaire. Forms should be designed so that they fit the type of information available. For example, if the available records do not permit differentiation between wholesale and retail trade, the questionnaire will have to cover both activities. Thus, the number of different forms and their complexity become a function of the information available for mailing. A special problem occurs when the file used for mailing forms includes some firms for which no classification is available. For such cases, a special form will need to be designed which will cover not only establishments in scope but, also identify the ones out of scope. However, data should be requested only for those establishments that are in scope. It may also be necessary to make a supplementary mailing to those establishments from which data obtained in the first inquiry are inadequate; this will be the case if the form designed for the initial contact is too long

and complicated by the inclusion of special inquiries pertinent to all kinds of business. The second contact can be limited to those establishments large enough to have a significant impact on the data. For smaller establishments, missing data can be imputed.

Design of report forms

127. Data collection forms are the key documents in surveys for recording all the data that are to be processed in preparation for tabulation and publication of the survey findings. Their format, organization and content will have a significant influence on the quality of the statistical results, the cost of collection, editing and tabulating and the promptness of publication. Ambiguous phrasing, printing errors and even poor spacing have caused some countries to abandon important inquiries or to expend scarce resources in an effort to correct the effects of such mistakes. In self-enumeration, when the respondent fills out a form without the assistance of an interviewer, even a poor selection of typeface has been known to cause widespread misinterpretation. This can also happen when a key heading, important for understanding the captions below, does not stand out boldly and is consequently overlooked by the respondent. Pre-testing of the data collection forms can often bring such deficiencies to light in time to correct them. The pre-test need not be elaborate or scientifically designed. If alternative formats are proposed, pre-tests can establish which format is preferable.

128. Proper use of space is also important in good form design. Too frequently, the space provided for the response to a question is inadequate, leading the respondent to feel that the surveying agency does not know its business. On the other hand, too much space makes the questionnaire larger than necessary, and the respondent may feel that there is more work involved in completing it than is actually the case. If it is expected that many questionnaires will be filled out on typewriters, the spaces between the lines should correspond to typewriter

spacing. In addition, the questionnaire should be examined considering space needed for office notations. Entries "for office use only", are of three general types:

- (a) Control data, such as establishment number, identification of those processing the reports and dates on which various actions take place;
- (b) Code numbers, such as those for location, kind of activity, size of establishment and type of ownership. To the extent that questions are designed to be pre-coded, this will not be necessary;
- (c) Calculated figures, such as the average number of persons engaged, average earnings etc.

129. Because names or addresses frequently need to be changed, the address box should leave sufficient space for this purpose. In addition, each questionnaire should include adequate space for the respondent to enter remarks that are necessary to explain any of his response.

Sequence and phrasing of questions

130. Whenever possible, it is advisable to place those questions that are easiest to answer at the beginning of the form. If the respondent gets discouraged as soon as he starts to complete the questionnaire, he is more apt to set it aside. Similarly, questions should be arranged so that those that proceed naturally from one to another are placed together. Where a questionnaire includes some special questions for only a portion of those receiving it, those questions should be moved towards the end. Each item on the questionnaire should be numbered to facilitate reference to it in instructions and correspondence with the respondent and in the writing of office procedures for handling the reports. Where more than one type of report form is used, every attempt should be made to retain the same number for the same item (for example, payroll and employment), even if this means that on some

forms an item number will have to be included with no data requested but marked "not applicable to this form". This will greatly simplify the writing of editing and clerical procedures.

131. Although instructions are important and should be carefully written, even greater care should be accorded to the wording of the questions. A major factor in securing prompt, complete and reliable replies is the phrasing of the questions so that they are readily understood by respondents and field interviewers. Some respondents will fail to read separate instructions, but presumably all of them will read the questions before entering the replies. Every effort should be made, therefore, to keep the wording of the questions clear and concise and to avoid ambiguities. If two interpretations of a question are equally likely and only one is correct, half the replies will be unacceptable.

132. For some items the use of a multiple-choice answer in which the respondent merely checks the correct box is preferable to a question calling for the respondent to make an entry. Such questions can be used to obtain information concerning the class of customer, the method of distribution of products, the legal form of organization etc. They may also be used for obtaining quantitative data if replies in terms of size classes are acceptable. To make sure that no category is omitted, it is advisable to include an "all other" category, requiring an entry to be made by the respondent when he feels that the available check-boxes are not applicable. These entries can be reviewed by a clerk to determine the box that should have been checked. The value of this check-box type of question is enhanced by the fact that pre-assigned codes can be printed next to each box, thus avoiding a clerical coding operation.

133. Finally, it should be mentioned that the use of colour and shading can result in better response, punching and general processing by making the person handling the report highly conscious of where entries are to be made and aware of

the fact that a blank space means that a question has not been answered. Devices such as arrows and leaders are also useful in this respect. Setting up a separate space for each digit in a value inquiry can also be helpful, because it requires making a more careful entry by the respondent; this makes the form easier for the puncher or clerk to process. The proper design of such an item can also result in pre-rounding, which not only eliminates a clerical operation but lessens the chance of error introduced by the clerk in performing this function.

VII. DATA COLLECTION

General remarks

134. In deciding what system is to be used to collect the data, it is necessary to analyse the special problems relating to the universe to be covered, the types of information to be collected and the availability of funds, qualified personnel and special equipment. For the distributive trades and services there are several characteristics that should be kept in mind in designing surveys. In many such kinds of activities there is a large number of small establishments. Frequently, proprietors of such establishments are illiterate and have great difficulty understanding concepts, even when personal enumeration is used. In addition, they tend not to keep records and, therefore, data for the period covered - whether year, quarter or month - are estimates which are usually poorer the longer the time period involved. Frequently, such establishments are operated in homes or in non-recognizable places of business or are conducted by itinerants. They tend to ignore requests received by mail. Because of the large number of establishments involved, personal enumeration becomes costly.

135. It is also important to consider, in survey design, the fact that, especially in the retail trades, a large share of the activity may be conducted by establishments of multi-unit enterprises. Since such an enterprise usually maintains a central office where data for each of its establishments is accumulated, it would be preferable to obtain the required data from the central office rather than have interviewers enter each location (where some of the necessary information may not even be available) or mail separate report forms to each establishment. If a list of most of such enterprises can be drawn up early in the planning for the infrequent censuses, a special pre-canvass can be conducted to obtain a complete

listing of all locations (including ancillary locations) which should include sufficient classification information to permit the use of the correct reporting form at the time of the actual canvass. All reporting forms can then be sent to the central office. This has the advantage of allowing the enterprise to be made responsible for co-ordinating the completion of returns for its establishments and reconciling any conflicts in the reporting of data among individual establishments. In any case, the central office will frequently prefer to supervise the filing of questionnaires and to transmit them when completed. In the use of such a procedure, however, great care must be taken to establish controls that will ensure that reporting is not duplicated either by field enumerators or by the mailing of forms directly to the establishments. Interviewers must be provided a list of enterprises or specific establishments that are not to be enumerated, and mailing lists must be properly purged of duplicated. When surveys are conducted subsequent to the completion of a comprehensive census, listings of multi-unit enterprises can be derived from census records and used as a skip list by field enumerators.

136. Another factor that affects survey design is the wide variety of activities and types of establishments covered by the distributive trades and services. Both retail and wholesale establishments are characterized by the transfer of goods, but information on the class of customer is essential for determining kind of activity. Furthermore, a detailed breakdown of activity is determined by different combinations of the kinds of goods sold. In service trades, the range of activities is extremely broad and includes such things as providing a meal, preparing advertising copy, producing a motion picture, renting machinery, repairing a clock etc. This requires that the system be designed to select the correct reporting form in the case of a mail canvass or that the enumerator be thoroughly trained to determine

kind of activity properly. There must also be provision for detecting and correcting errors when incorrect forms have been used.

Methods of enumeration

137. There are three basic methods that can be used to obtain data needed in the surveys. The first is the use of field enumerators who visit each place of business. If the establishment is small and the questionnaire simple enough so that answers can be obtained during the visit, the enumerator can complete the questionnaire at that time. If for some reason the questionnaire cannot be completed at the time of the visit, either because the person who can supply the information is not available or because the data requested is too complex, the questionnaire can be left to be mailed in or to be picked up by the enumerator on a subsequent visit.

138. The second basic method is that of mailing questionnaires directly to the respondents. This method can be used only when it is possible to develop a reliable directory of business establishments and the postal service works efficiently. Questionnaires may be mailed back upon completion, or enumerators may be used to visit establishments to pick up completed questionnaires.

139. A third basic method is used when a country has developed a fairly sophisticated administrative records system in which data needed in the survey are collected in the course of administering various national programmes, such as income tax, social security, unemployment and so on, and a method of positively identifying the establishment is built into the administrative system.

140. In practice, the three basic methods can be combined, depending on the particular conditions prevailing at the time of the survey in the country involved. Thus, it may be possible to develop a directory of large firms to which report forms may be mailed directly and to use this mailing list as a skip-list for field enumerators who would be assigned to collecting all remaining questionnaires.

Field enumeration

141. For countries planning to conduct a comprehensive census of the distributive trades and services for the first time, it would be most useful to conduct a complete field canvass in which a record is established in the form of a listing for each business establishment located, whether in- or out-of-scope. Such a listing can be very valuable in establishing sampling frames. Complete field enumeration is also necessary when it is not possible to develop a reliable directory that can be used for mailing questionnaires. In the use of a field enumeration, it may also be possible to use an area sample for portions of the nation that have only slight business activity, provided that the mapping of such areas is comprehensive enough to permit proper identification. In any event, the use of field enumeration requires an intensive effort to set up a detailed set of maps which will be used to control assignments to enumerators.

142. While a field enumeration of distributive trades and services has many characteristics that are similar to those used in a population census, there are differences which make the surveys of distributive trades and services much more difficult and require a higher calibre of enumerator. Incidentally, it is possible to combine a population census with one for distributive trades and services and theoretically save a great deal of money on mapping, enumerator training and overhead costs. However, the resulting complexity of the operation and the differing time requirements for completion would result in greatly complicating the operation and might lead to the collapse of both censuses.

143. The field enumeration of distributive-trade and service establishments is quite complex and requires the setting up of very tight controls and thorough training of field supervisors and enumerators. While it is not essential to have personnel with business training, the calibre of people used in this work must be

such that they can understand problems of classification. If they cannot grasp the concept of scope or if they have poor number sense, they can create great difficulties. Time must be taken to train enumerators properly both in the mechanics of enumeration and in the objectives of the survey. They must be made aware of the consequences of the actions they take--or don't take. Time and again it has been found necessary to re-enumerate substantial portions of a census because instructions have been misunderstood. If there are restrictions on the funds available, it would be better to reduce the scope of the survey (and, therefore, the number of enumerators required) rather than the quality of the enumeration.

144. In setting up training manuals and in the conduct of training sessions, it is advisable to concentrate as much of the work as possible in the hands of a few people on the planning staff. In practice, however, because of the large number of people who must be trained and the frequently wide geographical areas involved, it is necessary to train intermediate groups who will then train the actual enumerators. This results in dilution and distortion of the original material and should be kept to a minimum. A good procedure is to have persons who will eventually become field supervisors trained first and actually conduct interviews supervised by planning staff personnel to test how well they have absorbed the training material. Then they, in turn, can use this same procedure to test the field enumerators. Classwork is no substitute for actual field-work in determining how well training has been absorbed. Even though training material may have been well prepared, it should never be assumed that merely giving such material to enumerators is sufficient to get a good job done. There must be an early review of the enumerator's actual work so that mistakes can be corrected before much damage can be done. Ideally, this review would consist of recanvassing an enumeration district by the field supervisor.

145. Some advantage, in terms of training, results when the same enumerators are used in recurring surveys, such as monthly and quarterly surveys. However, even in such cases, it is wise to develop a regular programme of supervision of actual enumeration practices from time to time. Obtaining a good enumerator staff for a comprehensive census can be difficult, since the period of employment is quite short (although longer than is the case in a population census). Since this phase of the work may absorb the bulk of the funds, there may be a tendency to pay lower wages which makes it harder to attract competent help. Where there is an annual programme but no monthly or quarterly programme, investment in enumerator training becomes a very expensive item and the carry-over of personnel from one survey to the next may be quite small, since there is a long period between surveys and enumerators tend to get other jobs.

146. In addition to courses of instruction, the enumerator should have with him at the time of enumeration a manual that covers all the basic material he needs to know to conduct a successful enumeration. All procedural matters affecting enumerators and the principal problems they may expect to encounter should be included in such a manual. The subjects covered in a typical manual are outlined below:

- (a) Rules and regulations. Legal obligations of enumerators to provide true returns and treat information given by respondents as confidential. Actions forbidden to enumerators, such as soliciting or selling;
- (b) Excerpts from the law or executive order authorizing the census;
- (c) Procedures. Organization of field staff. How to obtain supplies and transmit completed questionnaires. What to do if respondent wants questionnaires left with him for completion or for mailing to statistical agency. What to do if respondent refuses to supply information. How to report establishments that changed owners during the census year. How to prepare progress reports;

- (d) The skip-list, that is, a list of establishments the enumerator is not expected to visit. This list may include the establishments of multi-unit enterprises that are to be contacted directly by the central statistical agency;
- (e) Scope of the census. Definition of an establishment and types of establishments to be included. Types of establishments commonly encountered that are not to be included;
- (f) List of questionnaires and other forms used in the census and a brief description of the content, purpose and scope of each;
- (g) Detailed instructions for each question in the questionnaire;
- (h) Special instructions for particular types of establishments or industries. How to handle frequently encountered mixed activities (for example, manufacturing and wholesale trade retail and service trade, leased departments);
- (i) Multi-unit enterprises not on the skip-list. How to identify establishments or headquarters offices of multi-unit enterprises;
- (j) Alphabetic index to the contents of the manual.

147. In preparation for field enumeration, each section of the country will have been divided into enumeration districts. For surveys based on the use of area samples, the selected areas will have been similarly subdivided into enumeration districts. Subsequently, a map for each enumeration district will be prepared for use by the enumerator in canvassing the districts to which he has been assigned.

148. The enumerator is required to list information for each establishment by entering it in a listing book containing all the information needed for control of the enumeration procedure. For each establishment, a determination will have to be made as to whether it is in- or out-of-scope for the survey. For in-scope establishments, entering the establishment and conducting an interview are required. In order to reduce the cost of field enumeration, it would be advisable to instruct the enumerator not to list such obviously out-of-scope locations as doctors' or lawyers' offices, farms etc. However, the importance of entering and listing any establishment about which there is any doubt must be emphasized.

149. In order to control the enumeration, to determine whether the establishment is in-scope, to decide what report form is to be completed and to be able to identify the establishment properly, it is necessary to obtain the following information:

1. Name and address of the establishment
2. Sufficient classification information to determine whether the establishment is in-scope
3. Whether the establishment is part of a multi-unit enterprise for which reports will be obtained from a central office
4. Whether a questionnaire has been obtained, will be sent directly to the central statistical agency or should be picked up by the enumerator at a later date
5. Date the questionnaire was actually obtained
6. Name of the person interviewed
7. Identification number of the establishment, to be used for control.

150. Because enumeration districts are pre-designated, it is possible to set them up so that their boundaries never cross geographical divisions for which it is desired to tabulate or publish data. Thus, if the enumeration district number is known, all the establishments that have been assigned that number can be geographically coded into provinces, cities or other areas.

Establishing and using a directory

151. It would be ideal if, before conducting a survey, a directory of all the statistical units to be included in the survey, along with information on location kinds of activity and size, were available. This would define the universe to be enumerated; information on kinds of activity could be used for preparing proper questionnaires, information on physical location could be used for geographical allocation and information on size could be used in sampling and coverage decision. The directory would permit control of the enumeration, whether by mail or by field.

enumeration, and serve as a medium for imputing data for non-respondents. Since it is impossible to prepare such a comprehensive directory until after the completion of a survey, the goal of the survey manager is to develop as close an approximation as possible prior to the survey.

152. As a minimum, the directory should include the following information:

1. Name and physical location of each establishment
2. Mailing address, which may be different from physical location
3. Name and address of the central office or headquarters of establishments that are part of multi-unit enterprises
4. Kind of activity, description or code
5. Size code, which should normally be based on number of persons engaged but which also may be based on volume of sales or receipts
6. Source and date of information

153. Because of the typically large number of small establishments in the distributive trades and services and because of the large volume of turn-over in ownership, the establishing and maintaining of a complete directory would be very difficult and expensive. As a result, some countries establish a specific size cut-off and include only in the directory those establishments over a specific size. Where other devices, such as a complete field listing or the use of other sources of data, such as administrative or tax records, are used to cover the remainder of the universe, this procedure is acceptable. However, if the survey is to be based on a cut-off that includes only those establishments above the size cut-off, there will always be uncertainty as to whether all units are included. Of course, if a country has developed a method of maintaining an up-to-date register of all business firms for administrative or tax purposes, such a register can be used to develop the information needed for the directory.

154. The usefulness of a directory depends on the kinds of data it contains, its completeness and its accuracy. There are a number of different sources for setting

up a directory, most of which have some deficiencies. Each source must be examined carefully before being used and care taken to overcome any defects. At the same time, it may be necessary to sacrifice some degree of completeness or accuracy in order to keep the costs of setting up the directory at a reasonable level, but such decisions should be made consciously and some attempt made to measure and describe the deficiencies. The following are some of the sources for directories of distributive trades and services:

- (a) A complete field listing. This is probably the most expensive way to collect information needed for a directory. However, especially when a country is initiating an economic statistics programme, it is undoubtedly the most useful. Trained field enumerators can seek out each physically recognizable place of business and collect the necessary information by direct interview and observation. Aside from high cost, the defects include the fact that non-recognizable places of business will be omitted;
- (b) Government records. Records maintained by the Government for taxation purposes and for the administration of unemployment insurance, social security or other government programmes may be helpful in setting up a directory. Such records, however, need careful review to determine their completeness, suitability for census purposes and accuracy, since they are not designed primarily to serve economic survey needs. The records may not have any indication of size and, thus, would not be usable for the selection of samples or implementation of a pre-determined size cut-off. They may relate to enterprises rather than establishments, thus requiring a preliminary survey (often referred to as a pre-cavass)

to collect information needed for each establishment. Addresses may refer to mailing addresses that are different from physical locations. Records may not be current; they may include firms that have changed ownership or gone out of business and may not include firms that have recently become active. The assignment of kind of activity (assuming such information is in the record) may have been on the basis of rules of classification different from those being used in the survey. One major advantage of using government records is that they will include listings for non-recognizable places of business and also for places that enumerators may miss because they are difficult to locate or that they may erroneously omit from the listing because of failure to understand instructions;

- (c) Trade association or publication directories. Because directories set up by trade associations or trade publications are designed for specific purposes, certain defects are introduced which users of such directories must be aware of. In addition to the problems that derive from the use of government records, trade association records tend to be restricted to association members only. Frequently, however, membership is extended to include firms that service the specific industry. Thus, a listing that purports to consist of beauty shops can include wholesalers and manufacturers of beauty shop supplies. Similarly, because trade publications are inclined to build up as large a clientele as possible, their listings will include many firms that have only fringe relations to the trade involved;

(d) Other potential sources. These include telephone directories or special listings prepared by telephone companies and lists prepared by private directory companies, chambers of commerce etc. Each type has its own special characteristics which must be studied carefully before a decision is made on how to use it.

155. One of the major problems in setting up a directory for use in a survey of distributive trades and services is the fact that enterprises engaging in those activities frequently tend to operate at a number of different locations. In addition, there is the problem of enterprises engaged in more than one kind of activity. In order to make the directory most useful, it is necessary to try to identify such enterprises and to contact them prior to the actual taking of the survey in order to be able to set up a separate record for each location which will include its physical location, kind-of-activity classification and some size indicator. Such contact or pre-canvassing can be based on a selection of enterprises known to have such characteristics (where such information is available).

156. In general, the directory is set up using one record for each establishment. For multi-establishment enterprises, however, there should also be a record for the central office, and each establishment should be cross-referenced to the central office. The records can be set up in the form of a plain card for each establishment or enterprise, or a set of punch cards can be prepared which can be converted to computer tape in those countries that have access to such a system. A separate record for each unit permits maximum flexibility and the elimination of records for establishments or firms going out of business, the addition of new records for those going into business, the introduction of changes and the sorting of records alphabetically, numerically, by kind of activity and by area. For countries having

a small directory, maintaining a set of plain cards may be adequate; however, the use of punch cards permits rapid sorting and greater flexibility. Computer capability is even more rapid and flexible but should not be used unless the volume and complexity of the work warrant it.

157. A common problem with the use of all sources for setting up a directory is the interval between the time when the information is obtained and the time when the directory is used. Because of it, the information is often incomplete and does not reflect changes that have occurred in the interim. If the directory is to be used in subsequent surveys, there must be some provision for supplementing it with new firms and corrections reflecting business deaths and changes in ownership. It is necessary, therefore, to build into the system a method for incorporating these changes. Care must be taken to make sure that this does not result in duplication in cases of firms that have merely changed their names or of one firm that has acquired another.

158. As indicated above, the best source of a complete directory becomes available upon completion of a comprehensive census. In such a census, much effort has been expended in identifying and classifying each statistical and reporting unit, and information on size, location and other characteristics has been assembled. Such a directory derived from a census can be used in setting up reporting panels to be used in annual and more-frequent-than-annual surveys that are based on the drawing of efficient stratified samples. These can then be employed in conjunction with an area sample. Such a directory should probably be limited to the larger enterprises and establishments, with all others being represented by the area sample; the area sample can then be used to represent small establishments as well as births of new establishments. The list of large firms to be included in the survey should not be restricted to those that fall into the area sample segments, and

data for such establishments should be collected directly by mail regardless of where the establishments are located. The determination of what size cut-off to use in defining large establishments should be based on the optimum balance between the cost and problems involved in field enumeration and mail enumeration. A listing of large establishments (and multi-unit enterprises) should be prepared and sorted by area to be transmitted to field offices and used as a skip-list in the enumeration of the areas selected in the area sample.

159. Since a directory represents a panel of firms that were in existence at the time of its preparation and is only as good as the source from which it was derived and the quality of its preparation, there is constant need either to update it or supplement it to include new firms and establishments. The use of an area sample in conjunction with a directory serves such a function. If other sources are available for keeping a directory current, such as tax and administrative records, they can be used. However, the processing and transfer of such records on a timely basis is extremely difficult.

160. Once a directory has been established, some countries set up a programme of data collection from a panel of the larger units included in the directory. When such a panel is used, the minimum requirement is to establish a system in which period-to-period trends are based on reporting by identical units. The obvious defects of such a programme include the fact that what happens to large units may not be representative of what happens in the total economy and that the units which fail to report may have different trends from those which do report. While this type of programme is least expensive to operate, the data it yields could be quite misleading.

161. As previously stated, an initial comprehensive census including a field listing operation is the best source for establishing a directory. In subsequent censuses, the list of establishments identified in the previous census becomes the

basis for a new directory. The use of such a directory in a census is, however, somewhat different from its use in the annual and more-frequent-than-annual surveys. The directory derived from the previous census becomes a control device which is used basically to make sure that the larger establishments and enterprises have not been omitted from the current census or that they have not been classified differently or allocated to a different area. Prior to the enumeration, data from the previous census, for large establishments, can be sorted by area and transmitted to the appropriate field offices for use as a check of the current enumeration. Field enumerators should be required to account for establishments not located. Once the enumeration is completed, records should be sorted by location, kind and size of business, and data from the two census periods should be matched. Significant differences and omissions should be researched. Care must be taken, however, to make sure that lack of a match is not due to change of name or ownership.

162. When a directory is established for use in a mail rather than a field enumeration, its function is different. In a mail enumeration, the directory must be complete in itself, and this requires that it contain some record for each unit for which data will be tabulated. The first requirement, therefore, is that the sources used be comprehensive and either contain the data needed for the census or permit contact with the firm to obtain such data. Since the census is restricted to certain kinds of activity, the source must include kind-of-activity information as well as name and address. It would also be very useful to have information on size. To be most useful, the directory should also include a separate record for each establishment of multi-establishment enterprises. When a directory contains all the above elements, it can be used not only as a means of controlling the enumeration but also as a basis for imputing data for delinquents and non-respondents. From such a directory, counts of establishments by kind of activity, geographical location and

broad size class can actually be tabulated. While these counts will not be as accurate as those that can be obtained by a completed report form, they can be useful for various purposes. When the enumeration is conducted, it will be possible, in effect, to prepare a schedule for delinquent reporters and refusals by ascribing to such establishments the same characteristics as those developed by tabulating reported data for establishments of the same size and in the same kind of activity, located in the same general area.

163. Where a common identification system has been established in different governmental files for the same firm, it is also possible to transfer actual data (although usually quite limited) from such files to a directory. This may make it possible to relieve an entire class of small firms from reporting directly in the census. In such use of administrative records there are several elements that must be studied to make sure of the validity of the procedure. First, it must be possible to identify the statistical or reporting unit; secondly, the assignment of kind of activity must be based on the same classification system; thirdly, the data to be transferred must be defined in the same way as they will be in the survey; fourthly, the administrative record file must be up to date.

Collection and control of reports

164. In a survey that calls for a field visit to establishments within the scope of the survey, the information obtained during the listing operation can be expanded so that listing and report collection procedures can be merged. As indicated above, since distributive trades and services are characterized by a large number of small establishments, a slight expansion of the information obtained at time of listing can make the listing book serve as the data-collecting vehicle for those small establishments from which only basic data are required. For all other establishments, the enumerator can leave the appropriate questionnaire and either set a date to call on

to pick up the completed form or arrange to have it mailed to the field office out of which he operates. Upon completion of his "list enumeration" work, the enumerator turns his listing book in to the field office which then reviews the book in conjunction with the reports it has received, both those collected by the enumerator and those mailed directly to the office.

165. As part of the listing operation, some indicator of size should be obtained for each establishment. This indicator can be used as a basis for the imputation of data for non-respondents. It is also useful as an indication of how much effort is warranted in attempting to collect reports from delinquent reporters.

166. When establishments belonging to multi-establishment organizations are encountered by the enumerator, the name and address of the central office from which the data are to be reported should be obtained. Cross-reference cards should be prepared for such establishments and sent to the central statistical office which can then collate them with similar cards from the rest of the country to use as a check on reporting by multi-establishment firms.

167. If the listing operation has been properly conducted and controlled, the establishments listed that are determined to be within the scope of the survey become the universe for which data will be tabulated. Subsequent operations should be tied to this universe if the survey is to be kept under control.

168. For practical reasons, it is desirable to detect and correct errors in the reported data as quickly as possible after the questionnaire or listing books have been completed. If there is a delay in re-interviewing or questioning the respondent, the person who supplied the data originally may no longer be employed by the enterprise, the enterprise may have gone out of business or the records may have been discarded. On the other hand, it is ordinarily not feasible or desirable to conduct complete editing operations in the field offices. The field office staff should,

however, be able to perform a screening operation which involves a limited amount of checking of data. This might consist of seeing that the questionnaire is properly identified, that no figures are omitted and similar simple checks. A slightly more complicated but basic check would be to add up the reported costs (wages and salaries, materials etc.) to see that they do not exceed the reported value of sales or services. If they do, the report should be returned to the enumerator or given to another interviewer for verification or correction.

Such a field review also provides some control over the performance of individual enumerators. When questionnaires or listing books are found to be faulty, the enumerator responsible should be required to correct them, if time permits. An excessive number of faulty questionnaires or listing books may mean the enumerator is incompetent and should be dismissed.

169. After a complete up-to-date directory of establishments and other reporting units has been prepared, it can be used to select those units that will be included in the survey, whether based on complete enumeration or a selected sample. Units will be selected according to kind of activity and will be enumerated using questionnaires specifically designed for each major kind-of-activity group. If the survey is to be based on a mail enumeration, the address in the directory will be used for this purpose and the respondent will be requested to return the completed questionnaire to the central statistical office within a given period of time, where it will be checked in upon receipt. From time to time it will be necessary to conduct follow-up operations for those firms that are delinquent in returning their questionnaires. The prompt and efficient conduct of these operations requires the design of a tight set of controls, strictly adhered to. Otherwise, there is a strong possibility of omitting some reports and duplicating others.

170. In addition to the proper questionnaire based on the kind of activity of the

firm, a number of other things must be provided in the mailing package. These include a self-addressed postage-free envelope which can be used by the respondent only for returning the completed questionnaire, a copy of the questionnaire which the respondent may keep for his own records and any special instructions needed by the respondent to properly complete and return the questionnaire. A statement on the confidentiality with which the reported data will be treated and on the penalties for non-reporting or wilful false reporting should also be included. Distinctively marked envelopes will draw the attention of the respondent to the importance of the contents. The use of window envelopes through which the imprinted address on the questionnaire can be seen avoids special addressing of envelopes. The date by which completed reports must be returned should be clearly stated. It is important to establish a deadline that allows a reasonable amount of time for completing the questionnaires but does not allow so much time that respondents set them aside and forget them. Generally, one month may be considered a reasonable time. It will not be possible to set a specific day and month for return unless all questionnaires are ready to be mailed to respondents on approximately the same date. If this is not possible, then respondents may be asked to return reports 30 days after receipt. Similarly, countries which have an unreliable mail system may have difficulty in establishing a specific date for return of the report and may have to request completion and return within 30 days after receipt. Under such circumstances, follow-up of delinquents becomes very difficult.

171. To make sure that each establishment included on the mailing list is accounted for so that the final tabulations may represent the entire universe, a tight check-in system must be established. This begins by the creation of a record, whether on regular cards, punch cards or computer tape, for each unit for which a questionnaire has been mailed. The assignment of a unique identification number to each such unit

will be helpful, since check-in by trade name can be complicated because of name changes, duplicate names, misspellings, changes in ownership etc. This identification number can also be used in the eventual filing of reports and will simplify finding report forms if they are needed during the course of tabulation review. Two types of listings will be found very useful during the processing of the survey: one in numerical sequence and the other in alphabetical sequence. While the use of a set of cards provides flexibility, unfortunately cards sometimes get lost or misfiled. Listings do not have this defect, although care should be taken that they are assembled in such a manner as to avoid mishandling, preferably in bound-book format.

172. As completed reports are received from respondents, a record is made which will permit culling the complete file at the time it is decided that it is necessary to follow up delinquent respondents. The follow up effort needed to achieve acceptable response rates is great, even in the more advanced countries. With a due date of one month after the questionnaires are mailed, it can be expected that no more than half of the reports will be filed within that time period. Because it would not be possible to tabulate the data without the inclusion of the largest units, it is advisable to establish special procedures for them. Such units should be visited prior to the survey to set up any special procedures that may be needed. During the survey, special efforts should be made to monitor the filing of their reports to make sure that they are received adequately in advance and that acceptable data are reported. The preparation of reports by such units may take more time, and the due date may have to be extended for them. Provision should also be made for a reasonable extension of due dates for units with special problems. Where an accounting firm has the responsibility of filing a large number of reports, it should also be allowed more time.

173. The collection of reports by mail will generate a large volume of correspondence, with respondents frequently raising questions that need answers before they complete the report. A special procedure should be designed to sort out such letters for prompt reply.

174. In deciding which system to use for data collection, it is important to keep in mind the fact that one of the main purposes of such surveys is to develop time series of data. Long-term cyclical trends are discovered from a study of census surveys; current patterns are traced in the more-frequent-than-annual surveys. In both types of surveys, however, the validity of the changes can be invalidated by a change in the method of conducting the survey. In designing the survey, therefore, the system that is adopted should be one that can be reproduced from survey to survey. If, for example, it is decided to use administrative records from a specific government agency in the setting up of a directory, there must be some assurance that the agency will continue to maintain the source material on an acceptable basis for the foreseeable future. Similarly, in deciding the extent and nature of the sample, there must be some assurance of the continuity of the process. Any "improvement" must be examined to see whether it results in a discontinuity of the data. However, if it is decided that it is necessary to adopt a procedure that will result in a discontinuity, every effort should be made to incorporate a "bridge" between the old and the new series so that users of the data can be aware of the effect of the change and make any necessary adjustments.

Use of sampling ^{6/}

175. In taking a survey, whether it be a comprehensive census containing a great deal of data or an annual or more-frequent-than-annual survey designed to produce

^{6/} A detailed discussion of sampling techniques is given in annex I.

fewer data as rapidly as possible, the survey manager will attempt to design his survey so that the required data will be produced at a minimum cost. Since sampling is one method by which costs can be reduced, it is advantageous to review each programme to determine to what extent and in what manner this technique can be used. In making a judgement, the manager must weigh the advantages derived from sampling, such as a reduction in the number of contacts to be made, the number of questionnaires to be processed and the manpower requirements, and earlier publication of data, against the disadvantages, such as the difficulties inherent in controlling samples, the higher quality of personnel required and the loss of data. In designing surveys in the distributive trades and services, the determination of the extent and nature of the samples to be used are conditioned by the heterogeneity of the kinds of activities to be covered. For example, a sample that might be good for measuring retail trade activity would probably be inappropriate for measuring advertising agency activity. A sample design that would be acceptable for a comprehensive census might not work well for a monthly survey.

176. When the survey procedure includes a field listing, it is possible to establish a size criterion above which all establishments are enumerated and below which a sample of every n-th case is selected for enumeration. This selection can be based on pre-designated lines in a listing book or can be selected after the listing books have been processed in the central statistical office. When properly inflated, these cases will then represent all establishments below the cut-off. Such data will, of course, contain sampling error. Instead of selecting only every n-th case below the cut-off size for enumeration, an alternative procedure would be to collect full information for such cases and have all other cases below the cut-off complete a very simplified questionnaire containing only the data. All establishments above the cut-off would be required to supply

complete information.

177. Where a mail enumeration procedure employs a directory that contains information on size and kind of activity, it is possible to select a stratified sample based on these characteristics. In designing any sample, however, it is important that large establishments be included with certainty. Because of the large number of small establishments, it is possible to design a sample that will considerably reduce the number of units to be canvassed, but the use of such a sample will considerably reduce the ability to publish area and kind-of-activity information.

178. The data objectives of annual and more-frequent-than-annual surveys are usually quite limited. In addition, fast production of the data is of great importance, especially in the more-frequent-than-annual surveys, since the objective is to collect information concerning current trends which will be used as a guide to early action. This is particularly true for items such as sales by retailers. As a consequence, the survey design should be as simple as possible and subject to rapid processing. In order to have validity, it should be a probability sample with known probability of selection so that sampling variability can be calculated. The most common design consists of the preparation of a list of large firms which will be canvassed each survey period and from which a system for selecting a sample of all other firms can be made. This sample can be either a sample of areas or, if a complete current listing of business firms is available, a sample selected from such a list. Since it is almost impossible to keep such a list current, this design should be supplemented by the use of an area sample which can be considerably smaller than would be required if such a list were not available. While the basic function of the area sample is to represent small establishments, it also serves the important purpose of finding and enumerating new establishments, regardless of size, that have come into existence subsequent to the selection of the "large" stores in the sample.

179. There are some kinds of activity in which the use of an area sample is quite inefficient and results in such large variances as to limit or even to rule out its use. This stems from the fact that the main use of the survey is for the development of data on trends in retail trade, since this is one of the most sensitive economic indicators. As a result, the selection of sample segments is keyed to population concentrations that correlate closely to retail trade activity. The fact that activities such as seasonal hotels, petroleum bulk stations, scrap wholesalers, motion picture producers and radio broadcasting stations which are also included in the scope are not typically located in areas that correlate with population concentrations means that the portion of the data collected in the sample segments will contain such high sampling variances as to make them useless. For such activities, therefore, it is necessary to rely basically on a fairly complete directory which can be kept up to date. If this is not possible then such activities should not be included in current surveys.

180. Theoretically, because the size of a sample means that fewer reports need to be collected and processed, its use has a number of advantages. These include relieving many respondents of the burden of reporting and by reducing correspondingly the workload of the statistical agency. This lower workload permits a greater degree of control to be exercised over the collection and processing of reports and, therefore, results in an improvement in the quality of the statistics produced. In addition, the timeliness of the data production can be improved, and it becomes possible to adjust the size of the sample to meet the availability of funds. Where funds are inadequate for a complete census, sampling may be the only possible method.

181. The sampling system used should be a probability sample which will yield results of measurable quality. The use of such a sample introduces additional

steps which must be properly executed. These involve some cost and create opportunities for mistakes that do not occur with complete coverage. Unless the sample is designed to fit the prevailing operating conditions and is satisfactorily controlled, losses rather than gains may result from its use. A highly elaborate and theoretically desirable sample design that requires more skilled personnel, more records or more equipment than is available would be worthless, or worse. Another drawback of sampling is that the amount of reliable detail that can be published will inevitably be reduced, because the smaller number of observations available for each cell of data results in an increase in the sampling variance, in many cases to such a level as to make the resulting statistics completely unreliable. Furthermore, even though collection, editing and tabulating can be better controlled with the smaller number of reports to be processed, errors in these functions cannot be completely eliminated.

VIII. DATA PROCESSING

General remarks

182. After the questionnaires have been received in the office, it will be necessary to subject them to a review which assures that they contain all the information needed and that they are processable in later operations. The review will include making sure that necessary data have been reported, that clerical rounding and coding are performed as required, that entries are legible, that write-ins by respondents are reviewed for necessary action and that clerical editing checks to detect erroneous data have been performed. This review will uncover a large number of problems which must be resolved before the report is ready for further processing. A high percentage of respondents will need to be contacted for verification or correction of data. Since this is very time-consuming, it should be initiated early and pushed vigorously so as not to delay later operations.

183. The most difficult function is the detection of errors in reporting. These errors may be due to various causes: respondents may not understand the concepts that are implicit for correct reporting; persons completing the reports may be careless; data may not be available in the units requested; data reported for employment and payroll, for example, may have different reference periods etc. These problems need to be detected and resolved prior to tabulation. The amount of work to be performed by clerical and professional staffs will vary, dependent upon the quality of personnel and the equipment available for this work. With the use of computers, sophisticated computations designed to detect errors can be performed at rapid rates. The less sophisticated the equipment available, the more work will have to be performed by clerical and professional staff.

184. It is a waste of resources to treat all deficiencies in questionnaires alike, and some standards and criteria for editing should be established. In general, the amount of effort expended on detecting and correcting errors and omissions should be in proportion to their probable effect on the published data. There is a certain amount of safety in numbers; if errors are of a random type, they will tend to cancel each other out, provided they are individually small. This leads to the conclusion that a high degree of accuracy and rigid standards are unnecessary when dealing with the small establishments, typical of many retail and service trades. The amount of attention paid to individual questionnaires during editing and coding should vary in direct proportion to the size of the establishments represented. It follows that reports from large establishments should normally be more carefully edited than those from small establishments, except where the coverage of small establishments is based on a sample.

Clerical operations

185. Assigning work, keeping records and progress reporting will be simplified if the questionnaires are sorted by type and bundled in work units of convenient size. Usually not more than 100 questionnaires should be assembled in a work unit; the actual number should be based on the amount of work which needs to be done but should not be so much as to tie up any work unit for an extended period of time. Work units should be numbered serially and a centralized control record kept giving a running record of each one, including the date assigned, the name of the clerk to whom assigned, the number of questionnaires requiring recontact with the respondent, the number of questionnaires referred for professional review and the date completed. A work unit usually consists of a single type of questionnaire. It may be advisable to sort questionnaires in other ways, such

as reports for small establishments separate from reports for large establishments, each of which may have its own set of editing rules. Similarly, reports for multi-establishment enterprises may be set up in a separate series of work units.

186. One of the major decisions to be made at this stage of the work concerns the degree to which screening and editing should be divided into highly specialized operations; each clerk can either repeat a few very simple routines that are learned quickly or can learn and perform all the steps necessary for the complete preparation of the questionnaire for punching. The first alternative has the disadvantage of requiring a rather complex control operation and much handling of the reports. The second requires the time to train a person who has the qualifications to handle more complex and difficult procedures. The decision will depend on good judgement and the type of personnel available. It may be advisable to strike a balance between the two extremes.

187. The clerical operation lends itself well to the use of a quality-control system designed to implement an early warning programme which would identify those clerks who are not following procedures correctly. The system should also be designed to measure production rate in conjunction with quality. Clerks should be notified immediately of their mistakes and those who continue to fall below acceptable standards should be dismissed. Frequently, this kind of check may also reveal an error in the procedures which could affect all editors.

188. It is the responsibility of the planning staff to prepare specifications which will be used by the processing staff to prepare instructions for the clerks on what tests should be made to determine whether the report is acceptable or defective. This staff should also supply the reference material that will permit the imputation of missing information and the coding schemes that need to be applied to the data. As with all procedures, the members of the planning staff

should review the results obtained. Staff members should arrange to check through some of the early work units immediately after their completion and to make any necessary modifications in the procedure as soon as possible. It is also advisable for them to visit clerical operations from time to time to see if any new problems are developing.

189. A correspondence unit should be set up as part of the screening and editing operations. It will be found that many of the answers that are needed to make a questionnaire complete or acceptable can be handled on the basis of a response to a form letter. The clerk can indicate the form letter to be sent (sometimes a single form letter can be designed to include a number of different items) and send a referral slip with this information to the correspondence unit which can dispatch the letter and set up controls to handle replies.

190. In some cases a simple procedure cannot be incorporated into the clerical processing operation; certain situations will occur for which a more complicated procedure is necessary. There should be set up a second group of more highly qualified clerks who have been specially trained to handle such operations. Problems that are unique and cannot be handled by either set of clerks should be referred to the professional staff for handling. Members of this staff will have the option of working out a solution themselves or of recontacting the respondent to solve the problem.

191. Among the clerical procedural rules that have merit and are used by many countries are the following:

- (a) Entries on the questionnaire that are made by editors should be made in a distinctively coloured pencil or ball-point pen. The colour chosen should contrast with that used in entries made by respondents or other persons handling the report;

- (b) Entries made on the basis of a respondent's reply to an inquiry resulting from editing should also be made in a distinctive colour;
- (c) When a correction is made to a respondent's entry, the original entry should not be erased or obliterated. The editing clerk should strike a line through the original entry in such a manner as to leave it legible and enter the corrected figure above it. If space is limited, the corrected figure may be entered elsewhere and its proper position indicated by a guideline;
- (d) A specific space should be provided on the form for the initials of the individuals (editors, coders, professional staff members) who have processed the forms and the date processed.

Preparation of the questionnaire for tabulation

Testing and checking

192. The first step is to determine whether the questionnaire covers an establishment that is within the scope of the distributive trades and services. This is especially necessary when the establishment has been contacted by mail and its kind of activity is in doubt. If ruled out of scope, no further processing will be necessary. Since testing for scope involves an inspection of the kind of activity engaged in by the establishment, it might seem logical at this point to also assign a detailed kind-of-activity code. The application of the rules involved in assigning such codes, however, can be quite complicated, and it would be advisable to reserve this operation for a group of clerks specially trained in this function. To assign activity codes, it will be necessary to use such information as sales by class of customer, volume of receipts from the performance of services as contrasted to volume of receipts from the sale of merchandise, sales by line of merchandise, services by type of service rendered, a self=

designated kind of activity and a historical code (when available). Rules on how to use this information to arrive at the kind-of-activity classification will be supplied by the planning staff which should also make its personnel available for the resolution of the many problems that might develop.

193. There are several items that must be present for every establishment included in the universe covered by the survey. These usually include total volume of sales, total employment and total annual payroll. (In more-frequent-than-annual surveys, the necessary data may be restricted to volume of sales and may include inventories in the case of retail and wholesale trades). Where needed data have not been supplied by large establishments, it will be necessary to recontact them. For small establishments, however, it would be more advisable to impute the missing data. A set of tables should be prepared to provide the basis for each imputation.

194. A particularly disturbing situation occurs when a response is requested to a question on an activity in which not all respondents may engage. Lack of response to such a question is ambiguous, since it could mean that the respondent either does not engage in that activity or merely hasn't bothered to supply the information. The tabulation of responses to this item might be misleading if the experience of those who did respond is different from that of those who did not, especially if omission of an entry indicated no activity for the item. If it is particularly important to distinguish failure to report from inapplicability, the question should include a "yes-no" feature as illustrated below:

Did you have income from _____ during the census year? Yes No

If your answer is "yes", how much was received from this activity during the census year? _____

If "yes" is checked but the amount omitted, it would be clear that the respondent had failed to report a figure that should have been reported.

Editing and coding

195. Frequently data entries turn out to be incorrect, and critical editing of the reported data is required to detect problem cases. There are two types of editing that should be performed. The first type checks the internal consistency of the data reported in different items for the same establishment, and the second checks whether the data reported for a specific item is reasonable on its own merits. Internal-consistency checks include checking to determine whether the reported components are equal to the reported total, whether answers to the same question in different parts of the report are the same and whether certain ratios, such as payroll per employee or payroll to sales, fall within acceptable limits. Determination of acceptable limits for the ratios being examined can be based on averages derived from data reported by similar establishments in the same industry and in the same geographical area. If similar surveys have been conducted in previous periods, the data derived from them can be used as a basis for setting up acceptable limits. Such data, however, should be adjusted for known changes in price levels and payrolls. In initial surveys it may be necessary to use data from other countries as a guide. One example of internal inconsistency is an answer of "zero" for receipts from the sale of groceries in an establishment that claimed to be a food store.

196. Other types of checks are needed to verify that the answer is reasonable. It would be safe to question an entry of "120 per cent" in response to a query on percentage of occupancy of hotels, television repair shops with no receipts for services, motion picture theatres with a seating capacity of 10, laundries with no laundering equipment etc. During the course of editing, the effectiveness of the

various checks should be examined in terms of the number of rejects being generated. If the majority of reports are being rejected as the result of the application of a given editing check, tolerances are almost certainly too narrow. If, on the other hand, practically no reports are being rejected, tolerances are probably too broad and appropriate changes should be made.

197. Depending on the extent to which the questionnaire has been designed so as to provide for pre-printed codes, the clerk will have to review the reported information and assign a numerical code that can be punched and used in sorting and tabulating. Even when codes have been pre-designated on the report form, it will be necessary to review certain cases such as the "write-in" answer which must be coded or the indication of two or more mutually exclusive alternatives.

198. The two most difficult coding operations are those for kind of activity and location. Determination of the correct kind-of-activity code requires that appropriate information be requested and supplied in the questionnaire and that frequently complicated rules be applied. Location coding may appear to be a simple matter of applying a numerical code to a place name, but experience shows that it is more difficult than that, especially in coding cities. Legal boundaries are difficult to identify, and people sometimes use city names in their addresses when the establishments are actually located outside the city boundaries. In a field enumeration, this difficulty can be overcome by having the enumerator determine establishment location; this is not possible in a mail survey. Other coding includes such items as legal organization, form of ownership, method of selling, merchandise line, various size codes etc.

199. Values are frequently reported in terms of the smallest monetary unit, but it is often not desirable to punch or tabulate the data in that manner. Punching the full number would require additional time and lead to a higher number of

errors; tabulating would also require more time and space. Similar problems occur when quantities are reported down to the units digit. Part of the preparation of the report for tabulation consists of reducing the number of digits to be processed by rounding the number. Training for this operation should be careful; many people do not seem to be able to grasp the principles involved. It is easier, of course, to use a procedure that calls for striking digits rather than rounding, but the data involved must be examined to make sure that this does not have a significant effect on the resulting statistics.

200. The report form should be reviewed for any notes or comments entered by the respondent which may require correspondence. As a final step, the clerk needs to review the report to make sure that the data to be punched are legible and that the puncher will have no difficulty in rapidly locating the items to be punched.

201. A survey that features extensive use of the mail to distribute and collect reports will generate a great deal of correspondence. Much of this correspondence is routine in nature and can be handled by the use of form letters. However, there will still be a large volume of work remaining for the handling of technical questions which will need to be resolved before the respondent can complete his report. If an adequately trained staff of people prepared to handle such matters is not assigned to this task, replies can easily be delayed for many weeks. This will not only delay completion of the survey but will frequently discourage the respondent from completing his report.

Computerized editing and coding

202. If the volume of work to be performed warrants it and if the proper type of personnel and equipment is available, much of the editing work can be performed by computers. But it will still be necessary for clerks to screen reports to make sure that key data items have been reported, that there are no obvious

errors and that any necessary work that cannot be performed by the computer is accomplished before the report is sent for punching. After punch cards are taped, the computer can perform many operations, such as computing ratios, testing limits, imputing data and assigning codes. It can also apply a series of tests to determine acceptability and identify cases that will require checking into or, in appropriate circumstances, substitute a computed value for an obviously erroneous one. In general, the computer can perform these operations more rapidly and accurately than clerks.

203. It is commonly believed that the use of computers will shorten the time required to produce final publications. However, experience does not justify this belief. Another common misconception is that use of the computer will automatically lead to a sharp reduction in staff requirements. This, also, is not necessarily so. There are a number of things that must be considered before a decision is made to use computers for editing:

- (a) Conversion of the planner's specifications to the machine language used by the programmer sets up an additional communications barrier. Unless the planner can read the programmer's language to make sure that his ideas have been correctly interpreted, there will be room for the introduction of erroneous matter. Planners will need to receive sufficient training to enable them to intelligently review the programmer's work;
- (b) Many of the reports that are flagged by the computer for manual review will require correction. It is costly to correct punch cards and computer tapes. Hence, if a great many reports will eventually require manual handling, repunching of cards and so on, computerized editing may result in a much higher over-all cost than clerical editing;

- (c) The particular computer that is to be used may have too little storage capacity to perform all of the editing checks efficiently in addition to other computerized operations, such as coding and tabulation. The programme should be tested as much as is necessary to determine the effect of storage limitations;
- (d) A computer with sufficient capacity is capable of far more checking than is feasible in a clerical operation. The computer may be able to perform several times as many checks for internal consistency, for example. Some planners, enthusiastic over this extraordinary capability, have specified so many editing checks that practically every report has been flagged for manual review. As in clerical editing, the usefulness of every individual check should be considered in relation to its cost and also the over-all cumulative effect of a large number of checks;
- (e) Computerized editing is necessarily performed at a later stage than clerical editing. As mentioned previously, it is desirable to question respondents as early as possible after the reports are received. The delay in recontacting respondents which is caused by computer editing must be taken into account when this system is used;
- (f) An important difference between computer processing and the older methods, such as hand and mechanical or electro-mechanical processing, is that with the former, misinterpretations giving rise to repetitive or systematic errors are often far more difficult to correct. Many costly tabulations have had to be scrapped because the computer programmer did not quite understand what the

planner wanted. Under the older processing methods, such mistakes were usually discovered and rectified before much harm had been done. The tremendous speed of computerized processing, however, means that entire projects are often completed before the systematic errors are discovered.

204. It is extremely important that sufficient time be allowed for the thorough testing of computer programs. A very useful device for testing programs is to tape a "test-deck" of punch cards similar to those that will eventually be processed but which contain errors and other material deliberately introduced to bring all the features of the program into operation. In this manner, the program can be tested at relatively little cost and the results used to correct the program where necessary. In practice, the production of a test-deck that will encompass all possibilities is often not possible.

205. The use of the computer has greatly reduced the amount of coding required, especially the amount that needs to be done clerically. In the first place, codes whose primary purpose is to facilitate the sorting of punch cards are not required. The computer can operate about as well on a data item such as number of employees as on a one-digit size code based on the original figure. Moreover, if a one-digit size code is required for other reasons, the computer can readily assign one. Secondly, clerks are notoriously weak when it comes to assigning codes based on a calculation and/or comparison of relative magnitudes, while the computer is always accurate in this respect, if it has been properly programmed. In general, the computer is preferred whenever a code number is to be assigned on the basis of numerical data, particularly if complex calculations are involved. Coding systems may be simplified also; while mechanical tabulation may be facilitated by a size= digit number in which each position is significant (for example, the first two

digits represent the province; the second two, the county; and the third two, the city), computer operations may be equally efficient using a three-digit number assigned serially to cities, with the reference to county and province kept in computer storage. Many of the practices geared to hand or mechanical processes need to be re-examined if a computer is to be used.

206. The assignment of kind-of-activity and location code by computer, however, involves some extremely difficult programming and is also quite costly, since it requires the use of a good deal of machine time and a computer that has enough capacity to store a large amount of reference material. These and other items involving non-numerical information can most often still be handled best by clerks.

Data from administrative records

207. The availability of information for business firms which is collected in the course of administering various governmental programmes presents an opportunity to use such information both as an original source of data and as an editing tool. Essential to such use, however, is an easy method of transferring data from one source to the other. One such method is to provide an identification number that is common to both the administrative records source and to the survey. While it is possible to conceive of transferring the administrative records data without the use of mechanical or electronic equipment, such a procedure is not very practical. The type of equipment that is most practical is the electronic computer which can sort, match and transfer data accurately at high rates of speed.

208. As indicated above, in order to make use of data in administrative records as a substitute for filing a report form, the administrative agency must use the same kind-of-activity classification, and the records must include the basic types of information needed in the survey, such as employment, payroll and volume of sales and any non-data characteristics needed, such as legal form of organization

and location. However, a secondary, but still important, use of the administrative records is as an editing tool and as a source of information for delinquent respondents. As an editing tool, the data from administrative records provide an independent method of checking on the accuracy of the data reported in the survey. For establishments that fail to file a report in the survey, the administrative record serves as a basis for imputing the needed information when other methods prove inadequate.

209. Because different agencies may pursue different data-processing techniques, it is frequently necessary to adjust the tape files to various requirements. This can be very expensive and time-consuming. As in all computer work, time and cost must be evaluated before a decision is made.

Editing in repetitive sample surveys

210. Because the operation of more-frequent-than-annual sample surveys typically involves the setting up of field offices strategically distributed throughout the country, the collection and editing of reports can be performed better by the field staff than a central statistical office. Business firms can be more readily contacted either in person or by telephone or mail to resolve problems. This is especially important, since there is usually great emphasis on speedy publication of data in these surveys, and answers to questions need to be obtained rapidly.

211. In those instances that involve the use of an enumerator to conduct an interview and to obtain the data at the time of the interview, the enumerator can be trained to perform some simple edit checks at the same time and to enter any comments needed to explain apparent discrepancies. These can be reviewed when his work is turned in to the field office. When the survey involves mailing the questionnaire to the field office, a series of checks can be performed there which should include not only a check for internal consistency but also a

comparison of data for the same firm as reported in previous periods. This is a very helpful feature of repetitive surveys, since the same firm will be reporting regularly over a long period, and a pattern can be established which will make it easy to detect errors.

212. The primary purpose of the field edit in more-frequent-than-annual surveys is to ensure that questionnaires are completely and correctly filled out before they are transmitted to the central statistical office which must process questionnaires from the entire country. At this stage, report forms are given only a limited review, because tabulations are produced under a tight time schedule. Any errors discovered after tabulation are difficult to correct.

General remarks

213. The types of tabulating equipment currently available range in complexity from simple desk adding machines to expensive, high-speed, large-capacity electronic computers with their specialized auxiliary equipment. When options are open to decide what equipment should be obtained, it is necessary to weigh a variety of factors, including the number of questionnaires to be processed, the quantity of data, the quantity and type of editing to be performed, the number and complexity of sorting and tabulating operations, the time constraints and the cost of alternative systems. The decision should be made by persons who are thoroughly familiar with all the types of equipment available and who also have a good understanding of the work to be performed. The theoretical advantages of electronic computers should be examined cautiously, and countries that have had experience in their use for similar work should be consulted before a decision is made. The major types of processing equipment are discussed below.

Adding machines

214. If the number of questionnaires is small and the intention is to produce sub-totals for only one or two characteristics of establishments, tabulation on adding machines may be faster and more efficient than other methods. As an example, suppose all items are to be aggregated by (a) kind of activity and (b) geographical area. After editing and coding, the questionnaires are sorted by geographical area and then by kind of activity. The data for each group of questionnaires representing an industry within an area are then summed on adding machines. If there are many items, the work will be facilitated by using a machine that records several columns at once. This kind of machine is sometimes

called a book-keeping machine. It should be of the type that prints out the data on paper for verification. After their accuracy has been verified, the totals obtained can be posted to summary work-sheets, and these can be summed again to produce (a) national totals by kind of activity and (b) area totals.

215. If an additional tabulation is required - for example, a table showing some or all of the items classified by size of establishment, the reports must be sorted again and the operations repeated. The amount of hand-sorting usually becomes too time-consuming if the data are to be tabulated by more than two characteristics, and therefore tabulation by other means should be considered. There are devices that facilitate sorting, such as cards with marginally punched holes on which the data are hand-posted or typed. These devices are sometimes useful in repetitive surveys but in infrequent inquiries are not likely to offer any advantages over adding-machine tabulation.

Electro-mechanical processing

216. The term "electro-mechanical" is used to describe the kind of equipment that involves the use of punched cards in the mechanical tabulation of data. The basic types of equipment needed in this system include card-punch machines, card-verifying machines, sorters, collators, reproducing punches and accounting machines (tabulators). With this type of equipment, data from the questionnaires are first transcribed on to punch cards by use of the card-punch machine, and the punching is verified by another operator using the card-verifying machine. Cards are then sorted and collated as needed for the particular tabulation or listing wanted; tabulations are prepared by use of the accounting machine.

217. Electro-mechanical tabulation equipment varies in speed and capacity. The speed depends in part on the number of sub-totals and totals required. Some models are capable of calculating and printing out minor, intermediate and major

totals for the same data field - for instance, data for provinces, cities and the nation as a whole. Some can produce new punch cards showing summary totals for selected categories. These summary cards can then be used in subsequent tabulations.

218. After data have been transcribed on to punch cards, the cards can be used in conjunction with electro-mechanical tabulators or as an input to electronic computers - either directly, when smaller computers are used, or indirectly, after their conversion to computer tape. The use of punch cards in conjunction with electro-mechanical sorters and tabulators is especially useful in economic censuses in which the basic data may be classified and cross-classified by various establishment characteristics, such as kind of activity, size, location and method of operation. When used in conjunction with electronic computers, even greater flexibility, speed and accuracy are possible.

219. Some punching machines are suitable for punching numerical data only, while others will also handle alphabetical information. Some types automatically print each character at the top of the card as it is being punched, and this can assist in the verification.

220. The punching of cards is an operation that cannot be performed without errors. These are frequently more significant than those that arise out of hand transcription. By failing to skip spaces properly, for example, the punch-card operator can easily transform the number 123 into the number 123,000. Such errors can do great harm. It is therefore essential that all punching go through a verification operation which consists of another operator punching the data on a verifying punch machine. In general, however, verification punching has proven to be only partially effective. In a verification operation of average efficiency, it may be expected that about 95 per cent of the errors will be detected.

221. Punching and verification will normally be faster and more efficient if the cards are designed so that the figures are punched in the sequence in which they appear in the questionnaire. In electro-mechanical processing, it is desirable to group on one card the items that are to be tabulated together. If, to facilitate punching, the items have not been punched in the proper sequence for tabulation, new cards can be created on a reproducing punch machine, but this will add to the cost. If there are two or more cards per questionnaire, the amount of identification and classification information required on each card is also affected by the method of tabulation. With electro-mechanical processing it is customary to reproduce on each card the establishment identification number, the kind-of-activity and geographical codes and all the other codes to be used for the classification of the data, such as size codes.

222. With the use of computers, data fields can often be regrouped at little or no cost as they are transferred from cards to magnetic tape. When key-to-tape equipment is used with computers, the full information need be keyed only once, which is not the case with electro-mechanical equipment. The computer can then electronically assemble the complete questionnaire for an establishment and rearrange it in any desired format.

223. Developments in technology have made it possible to avoid entirely the production of punch cards as a medium for the entry of data into the computer. The two major types of systems are those that use tape as a medium (key-to-tape) and those that have direct access to the computer (key-to-disk). There is also a system that uses an "electronic pen" in place of a punch machine to transfer data directly to a computer. While it is not necessary to first create a set of punch cards or a tape for access to a computer, it is possible, once the data have been entered in the computer, to prepare a tape of such data, and it is also possible to create a set of punch cards from a computer tape.

Computer processing

224. The development and use of electronic computers in the past 20 years have proceeded at a tremendous rate. The result has been a rapidly expanding technology in which new ideas have led to continuous changes in equipment and in operating procedures, with a marked effect on the field of data processing. However, these changes have also brought about much confusion as several successive "generations" of computers and auxiliary equipment and a number of computer languages have appeared on the scene. In several places in this manual, warnings have been made on the potential dangers of plunging into this method of data processing. This is not to gainsay the potential advantages of speed, accuracy and flexibility, but merely to caution that the computer can only follow instructions and that the writing of such instructions must be complete and accurate in every detail. Depending on the size of its memory and the type of auxiliary equipment used, the electronic computer can handle vast amounts of data, perform numerous sorts and calculations in response to complicated instructions and print out the results in the desired format. The tabulation of industrial census data on adding machines can conceivably be started on a few hours' notice, provided machine operators are available. The use of electro-mechanical equipment will require more time for planning and for wiring the control panels. In computerized tabulation, however, considerably more notice is necessary. First, a trained programmer must be found and he must take time to familiarize himself with the objectives of the census, particularly with the plans and specifications for editing and tabulation. This may involve intensive and extensive systems design before programming is undertaken. Next, he should plan the broad outlines of the computer program and set down its essential features in the form of a flow chart. At this stage, the programmer and the planning staff should discuss the chart in detail, taking as much time as

is necessary to make certain that it is completely understood by all concerned. The objective of this review is to bring to light misconceptions and misinterpretations which, if not detected, may be extremely costly to correct later on. The need for mutual education (that is to say, the programmer has to be aware of the statistician's problems and vice versa) and clear communication, which is the biggest problem in computerization, cannot be over-emphasized.

225. After approval of the general plan, the programmer will begin to write the program itself. The time needed for programming depends on the complexity of the program and the type of computer to be used. Simple programs may take only a few days to write; others may take months. After it is written, the program should be tested. This is usually done with a "test-deck" of punch cards of the type to be tabulated, which has been designed to bring out the features of the program and to represent the problems anticipated in dealing with the actual data. After testing and correction, the program is ready for use in editing or tabulating the reported data.

226. Automatic verification of the punching of coded information is made possible by means of the check digit - a number determined rapidly by the computer from the code number itself and then added as the last digit. With a six-digit code, for example, the check digit becomes the seventh digit. Check digits can be calculated in many ways. In a computerized editing operation, the check digit is recalculated, and if the result does not conform to the original result, the code is flagged as incorrect.

Review and correction of tabulations

227. The efficacy of the work that has been done in designing the data-collection system and processing the individual questionnaires can be tested in earnest when data are accumulated and presented in the form of tabulations. It is only at this

stage that a judgement can be made as to the reasonableness of the results. To make this judgement, it is necessary to have available a number of different tools that can be applied, by personnel who have been trained in their use, to the evaluation of the tabulations.

228. While the use of sophisticated equipment such as the electronic computer makes it possible to perform more complex data tests, there still appears to be no substitute for the individual who can look at a set of numbers and, in light of his training and background, determine whether they make sense. However, the analyst should never make any arbitrary adjustments on the assumption that, because certain data appear incorrect, they are wrong. Unusual situations sometimes do occur, and the data should be checked with the establishment supplying them. The review of final tabulations should not be an occasion for repeating the work performed in the earlier stages of processing. At this stage, data review consists of examining summary data to determine its reasonableness. Only if the data cell fails the given test should there be any attempt to determine whether the problem is caused by previous operations. An adequate budget must be established for this phase of the work, since failure to identify and correct the errors that result in unusable data will mean that all the effort put into prior operations has been wasted. The mishandling of one report can result in invalidating data for an entire industry.

229. There are two major types of error that can cause difficulty. The first type consists of the inclusion of incorrect data (whatever the reason) for one large establishment or a small number of them which is evidenced by a non-systematic failure of a data cell to pass an edit test. The second type of error is systematic and affects a large number of cells for a specific data item. This type of error may show up even when data for only small establishments are being reviewed and can create a great deal of difficulty, since it may require redoing an entire operation.

230. Tabulations prepared at this stage, while designed to contain all the information that will be presented in the final publication, need not be set up in publication format. Instead, they should be structured basically to facilitate analysis. The arrangement and spacing must be designed so as to simplify the reviewer's ability to isolate problem data cells and to supply as much assistance as possible in identifying the causes of the problem.

231. There are two types of checks that can be applied to the editing of cell totals: checks for internal consistency and checks against available external sources of information. The checks for internal consistency seek to isolate those cells of data that fall outside an acceptable range when one data field is related to another. In the analysis of tabulations, however, the acceptable ranges should be considerably narrower than they were for individual establishments inasmuch as we are now dealing with an accumulation of data for a group of establishments in which the effect of individual extreme cases is greatly modified. As a corollary, the larger the number of observations in a cell, the smaller should be the acceptable range and vice versa. A set of tables should be prepared outlining acceptable limits in terms of the number of observations in a cell. In determining acceptable ratios, provision should be made for such factors as variants due to geographical location and kind of activity. If a computer is used in preparing tabulations for review, this table can be programmed into the computer as a reference file, and cells that fall outside of the established ranges can be identified by the computer and flagged for special attention.

232. The tabulation that will be found most useful for detecting problems is the one that presents data in the greatest detail. This usually consists of a table providing data on a detailed breakdown of activity for the smallest identifiable area and presenting basic information on number of establishments, sales or receipts, employment and payroll. The detection and correction of significant errors at this

level of detail will, when carried to higher levels, result in the production of acceptable tabulations at all levels in so far as the same items of data are included in them. To the extent that additional items of data (such as, for example, number of guest rooms) do not appear in other tabulations, they will naturally have to be reviewed independently.

233. In addition to the performance of internal-consistency checks, data can be reviewed against information available from previous surveys and other sources. Some difficulty in making this kind of analysis will stem from changing price levels and wage ratios over time. These may vary considerably from one kind of activity to another, and the analyst must be aware of these differential rates of change.

234. Once a data cell is identified as a potential problem, the analyst must find ways to determine what, if anything, is wrong. If the problem appears to be a systematic one, he must be able to trace the operation back to the point that might have generated the difficulty. For this he needs to be familiar with all that has gone on prior to tabulation. If the problem is not systematic, he must somehow be able to identify those establishments within the data cell that could potentially cause it. This means that he needs to examine the original data source - the questionnaire - or its equivalent. To be accessible, the report forms need to be in the same sequence as the tabulation being reviewed, preferably by area, by kind of activity, by size. Since the problem will almost always be restricted to large establishments, the sorting of schedules can be limited to that group.

235. With the use of computers, it is possible to prepare listings of various types that can help to resolve many problems without the need to resort to the questionnaires. Data-processing machines can rapidly sort large files into the

most useful sequences and eliminate the need to refer to bulky report files in which forms are often lost or misplaced. Thus, listings can be made and bound in book form for ready reference for such sequences as area by kind of activity, kind of activity by size based on various characteristics, industry by area etc. Even with such listings, however, there may be occasional need for a review of the questionnaires.

236. Once the need to make a correction has been determined, the problem arises as to how best to introduce the correction. Where a data cell appears in only one place, this presents no great difficulty, but in tabulations of economic data, it is typical for the data of an establishment to appear at different levels of area and kind of activity and in cross-classifications. This makes it essential to be able to introduce the correction at many points in the tabulations. These corrections can be introduced by hand or they can be carried to the punch-card or tape files either at the individual establishment level or at the lowest summary-card level and all tabulations affected by the correction redone. If it is expected that the individual establishment records will be used for preparing additional tabulations in sequences other than those to be published in the survey, it will be necessary to correct such records, even though that would not be required for survey purposes in which only summary cards are used in preparing final tabulations.

237. There should be a single source document on which to indicate all the corrections that need to be carried out for the establishment or data cell. The entry of new or corrected data at this stage must be exact, inasmuch as it will be used in the preparation of the final publication tables, and the verification of the carrying of corrections must be absolute. There is no margin for error at this point.

238. Even with the use of electro-mechanical and electronic computer equipment, it is necessary to check to make sure that the sum of the detail shown in the vertical total column of a tabulation equals the sum of the detail shown in the column heads. It is surprising how many discrepancies may be detected. An investigation of the cause of such discrepancies can lead to the discovery of serious errors.

239. As in the case of the review of cell data for problems, the best time to search for instances in which the publication of data will, in effect, disclose information concerning the operation of a business establishment or enterprise is when data are reviewed in the greatest detail. At this stage, information must be sought on whether a single establishment dominates a cell of data (such as food stores in a city) to such an extent that publication of the data will permit competitors, government officials or others to have a very good idea of the income or other important characteristics of the unit. It is at this point that it becomes most useful for purposes of comparison to list the largest firms in the sequence in which they appear in the tabulations.

240. The usual approach to disclosure analysis is to suppress cells that, if published, would reveal the activities of a single statistical unit. In this situation, the unit is normally an establishment, but the rule is often extended to the activities of enterprises that own or control two or more establishments. The publication of data for a cell containing only two statistical units is prohibited, because the activity of each unit would be revealed to the other. Thus, publishable cells must include, as a minimum, data for three statistical units.

241. Some statistics are considered to be more confidential than others. The volume of receipts and amount of payroll are examples of data generally regarded as particularly sensitive. The number of establishments is generally regarded as non-confidential unless the classification is so detailed as to reveal confidential information.

X. PUBLICATION

Final review

242. Before data are typed or typeset (or before computer-generated publication tables are run for the final time), they should be subjected to a final review. For this work, only the most highly qualified persons should be used. While clerks and other personnel should be available to perform the necessary digging to extract the information required by the analyst, the determination of what items look suspicious and the decision as to what needs to be done should be made only by him. His aim is not to detect small errors or to double-check previous editing, coding and corrections but to examine the data in light of his own knowledge, experience and intuition. If it is discovered at this final stage that the quality of the statistics is poor, whether due to response or processing errors, or that, in the case of sample inquiries, the variance due to sampling is too great, it may be better to suppress the data rather than publish it.

243. In conducting the final review, the analyst should have, whenever possible, measures of the extent of the estimating and imputation that have gone into production of the final figures. In addition, he should be aware of any technical difficulties that have been encountered and how they were overcome, so that he may use such information not only for his own understanding of the statistics but also in preparing any needed explanatory foot-notes or text discussions of data limitations.

Final disclosure analysis

244. Direct analysis of data to uncover those cells the publication of which

would result in the actual or readily computable disclosure of information concerning establishment operation will have been accomplished during the review of tabulations. However, there is another aspect of disclosure analysis that should be undertaken at the time of final review. This is the search for complementary disclosure cells. In statistical tables with horizontal or vertical totals, a complication occurs when one of the components of a total is suppressed, for if at least one more additional data cell is not also suppressed, the original disclosure cell can be derived by subtracting the sum of the published detail cells from the total. The greater the amount of detail for which information is being tabulated, the greater will be the number of direct disclosure cells; when complementary disclosures are added, the result may be so many suppressions as to make the table virtually useless. In planning the level of publication detail, therefore, the fact that attempting to publish more detail may actually result in publishing less information should be taken into consideration.

Copy preparation and duplication

245. After tabulations have been reviewed and any necessary corrections made, the next step is to prepare tables in final form for release to the public. As discussed above in the section on publication planning, the basic work on table content and format should have been performed early in the planning stage, since so many of the survey operations depend on what will be published.

246. A key factor in determining the time required and the cost of publication preparation and duplication is the method of copy preparation and reproduction. There are three basic approaches. The first consists of the hand posting of tables, including the provision of stubs, column heads, titles, foot-notes etc., after which the copy is sent to a printer for typesetting, printing and binding. While this approach will result in a very impressive-looking publication, it has

the disadvantages of being very time-consuming and expensive. The second basic method also involves the hand-posting of tables which then are typed exactly as they will be published and reproduced by the photo-offset method. While this method may not produce as good-looking a publication, it has the virtue of being less expensive and faster to produce. The third basic method involves the preparation of final copy on the computer. While this is the fastest and potentially the least expensive, it requires extensive and difficult computer programming and should be undertaken only when the number of pages in a particular table is large enough to warrant the programming effort and computer cost. In this method, the copy would also be reproduced by the photo-offset method, although techniques are being perfected to use computer tape as a medium for typesetting. Computer printouts can also be used for hand-posting and typing in conjunction with either typesetting or offset copying.

247. No medium for copy preparation is infallible; all methods require careful review to make sure that errors do not creep in during the process of hand-posting, typing, typesetting or computer operations.

Publication of preliminary reports

248. The statistical agency will always be under great pressure to publish data as early as possible. This pressure may lead to the temptation to release information before its quality can be adequately verified. If errors that require substantial revisions are subsequently found, the agency will find itself subjected to a great deal of criticism. A compromise that, while not completely satisfactory, is still useful consists of the issuance of preliminary reports which contain considerably less detail than will be included in the final reports and are clearly marked as being subject to revision. If at all possible, the summary figures presented for advance release should represent control totals that have been confirmed and will be repeated in the final publications.

Inquiry descriptions to be published

249. The international recommendations adopted by the Statistical Commission request that in order to enable the users of published statistics to evaluate their comparability and assess their reliability, the following information should be published:

- (a) A description of the scope of the inquiry - that is, the activities that were included in the field of inquiry - and a definition of the statistical unit in terms of the field of inquiry and the items of data gathered;
- (b) A description of the coverage of the inquiry - that is, whether units of all kinds (size, type of ownership, kind of legal organization and so on) were included;
- (c) A description of the methods of covering the field of inquiry - that is, whether by direct collection (mail or field enumeration), administrative reports or sampling (including a description of the sample design used and estimates of probable sampling errors). This description might also include an evaluation of the completeness of the coverage attained;
- (d) The operational definitions of the items of data gathered, perhaps in the form of copies of the questionnaires and of the basic instructions, including a description of the valuation methods used for the various items. The definitions of derived items that have been compiled from the items of data gathered should also be given;
 - (i) The number and the importance of known units failing to respond to the questionnaire, together with, if possible, some of the key characteristics of these non-respondents - that is, their kind of activity and, particularly, size.

Also, whether estimates for these non-respondents have been included in the published data;

(ii) The extent of the non-response to particular questions for which no estimates have been or could be made;

(e) A description of the kind of activity and other classifications used. ^{7/}

250. In addition, it would be useful to include in the text an analytical review of the highlights of the survey findings. The review can be based on the preparation of a set of short summary tables which would feature such things as ratios (e.g., inventories to sales, payroll to sales, payroll per employee) by kind of activity and geographical region and percentage distributions (e.g., sales by sales size of establishment based on volume of sales, number of employees, geographical region). These could be supplemented by the preparation of charts which make the statistics more vivid and point up important aspects of the survey.

Content of tables

251. To be most useful, especially in the presentation of data for periodic censuses, the major tables should include some information for previous census periods. Such data can be included in the summary tables showing statistics by kind of activity and by area. It would also be helpful in such tables to show percentage change from the previous census to the current one.

252. Information concerning retail trade and personal services is most useful to establishments engaged in those activities in direct relation to the smallness of the area for which data are published. This is especially true for retail

^{7/} International Recommendations on Statistics of the Distributive Trades and Services, para. 100.

trade. Establishments are not so much interested in data in terms of the physical boundaries of cities and provinces. Despite pressure to produce data for major retail centres and central business districts, this type of work should not be undertaken without a clear picture of the amount of work, time and money it will require.

253. A special problem in tables for the service trades stems from the heterogeneous nature of those trades. It is further complicated by the fact that the recommendation on scope of the service trades, while "drawn up (by) selecting the activities most likely to be common to all countries", is qualified by the statement that "the services listed do not all rank equally for collection of data and it is difficult to specify separate priorities for each. For a particular country, adjustments may be required in the scope of the industries, such as the deletion of unimportant or non-existing services and the inclusion of others which are of some importance in the country concerned". It is obvious from these remarks that for international comparability, the totals for all service trades need to be carefully reviewed for scope from country to country. Also, the kinds of activity that are combined under this heading - real estate, motion-picture production, hotels, advertising services, radio broadcasting, laundries, data-processing services etc. - make the meaning of a combined total difficult to grasp. Basically, data for many of the service trades have meaning only when each industry is examined separately. This frequently calls for the collection of specialized data for the industry and the preparation of specialized tables. Where enough of this specialized information is collected, the publication of a special report for the industry may be warranted.

254. Tables containing data based in whole or in part on the use of samples should be clearly identified as such. Data should always be shown on an inflated

basis. The fact that totals based on the use of samples will not be identical with totals based on complete enumeration should be explained, and estimates of variance due to sampling should always be included.

255. Practically all tables for publication will consist of basic data items cross-classified by establishment characteristics. The number of tabulations that could be prepared in this way is so large that countries have to select the few that will be of the most general interest. Examples of some of the possible tables are suggested by the table titles and brief descriptions that are given below. The tables considered are based on the list of statistics to be tabulated according to kind of activity included in the international recommendations. The list is better suited for the tabulation of census results, since some items considered are not recommended for collection in annual or more frequent surveys. It is not intended to be comprehensive, and the titles are not necessarily in order of importance. Presentation of data according to geographical area, in addition to what is suggested for table 1, should be determined by individual countries in the light of their requirements and available resources.

Table 1. Principal indicators of activity, classified by ISIC group and by area

This table would present basic data items, such as number of establishments, number of persons engaged, number of employees, wages and salaries paid to employees, value of sales of goods and operating receipts, value added and value of gross fixed capital formation. The ISIC groups could be the ones listed in paragraph 62 of the international recommendations. The area classification would correspond to major administrative areas in the country,

such as states, provinces etc. If larger geographical detail is required, it would probably be necessary to reduce the level of the activity classification to the three-digit major group or two-digit division of ISIC.

Table 1a. Wholesale and retail trade: principal indicators of activity, classified by ISIC group and by area

This table would present, for wholesale and retail trade only, the same basic data items as table 1 plus gross margin, classified by ISIC group and by area. The ISIC groups could be the ones suggested in paragraph 62 of the international recommendations or a comparable national classification. As with table 1, individual countries should consider the level of detail of the geographical classification against the level of the activity classification so as to have a useful table of reasonable length.

Table 2. Principal indicators of activity, classified by ISIC division and by size of establishment

This table would present the same basic data items as table 1, cross-classified by kind of activity at the two-digit level of ISIC and by size of establishment. Obvious criteria for the definition of size are the number of persons engaged and the total value of sale and receipts.

Table 2a. Wholesale and retail trade: principal indicators of activity, classified by ISIC group and by size of establishment

This table would present, for wholesale and retail trade only, the same basic data items as table 1a, cross-classified by kind of activity at the four-digit level of ISIC groups suggested in paragraph 62 of the international recommendations or a comparable national classification and by size of establishment.

Table 3. Principal indicators of activity, classified by ISIC division and by type of establishment

This table would present the same basic data items as table 1, cross-classified by kind of activity at the two-digit level of ISIC and selected classes of economic and legal organization and ownership. If the number of classes for the latter characteristics is large, it would be preferable in practice to split this table into two or three parts, one for each characteristic.

Table 3a. Wholesale and retail trade: principal indicators of activity, classified by ISIC group and by type of establishment

This table would present, for wholesale and retail trade only, the same basic data items as table 1a, cross-classified by kind of activity at the four-digit ISIC groups suggested in paragraph 62 of the international recommendations or a comparable national classification and by type of economic and legal organization from among selected classes, type of operation and type of ownership. If the number of classes for the latter characteristics is large, it would be preferable in practice to split this table into two, three or four parts, one for each characteristic.

Table 4. Indicators of employment and earnings, classified by ISIC group

This table would present for the nation as a whole data items on employment and earnings, such as number of working proprietors and unpaid family workers, number of employees (if possible, distinguishing full-time employees), wages and salaries paid to employees (if possible, distinguishing payments to full-time employees) and supplements to wages and salaries. The data would be classified at the four-digit level of ISIC or the national equivalent.

Table 5a. Wholesale and retail trade: value of sales and receipts, classified by ISIC group

This table would present, for wholesale and retail trade only, data on the total gross output and its components, such as value of sales of goods classified in wholesale sales (further classified into sales on own account, on account of others and commissions received) and retail sales, and total operating receipts. Data would be classified at the four-digit level of ISIC groups suggested in paragraph 62 of the international recommendations.

Table 5b. Wholesale and retail trade: value of sales, classified by kind of commodity

This table would present, for wholesale and retail trade only, data on the value of sales of goods classified in wholesale and retail sales by kind of commodity or commodity groups. Each country needs to draw up special lists for the reporting of sales by kind of commodity. It would be useful if the list of retail commodities were linked to the classification of household goods and services for national accounting purposes. ^{8/}

Table 6. Value of purchases and other costs, classified by ISIC group

This table would present data on the value of input and its components, such as value of purchases of goods (those intended for sale and others), electricity and fuel purchased, repair and maintenance work, contract and commission work and other operating costs. Data would be classified at the four-digit level of ISIC or the national equivalent.

^{8/} Guidance on the development of such a classification will be found in the studies prepared by the Statistical Office of the United Nations in the course of developing a generalized commodity classification. See "International standard classification of all goods and services" (E/CN.3/493).

Table 6a. Wholesale and retail: value of purchases of goods for sale, classified by kind of commodity

This table would present, for wholesale and retail trade only, data on the value of purchases of goods for sale in wholesale trade and in retail trade by kind of commodity or commodity groups.

Table 7a. Wholesale and retail: value of stocks, classified by ISIC group

This table would present, for wholesale and retail trade only, the value of stocks at the beginning and end of the year, distinguishing stocks of goods for sale from other stocks, if possible. Other items to be shown in this table are changes of stocks during the year and stock turnover. The classification of the data would be at the four-digit level of ISIC groups suggested in paragraph 62 of the international recommendations.

Table 8. Gross fixed capital formation and its composition, classified by ISIC group

This table would present data on the value of acquisitions of fixed assets and gross additions to fixed assets, according to type of asset (that is, buildings and other construction, land, transport equipment, machinery and other equipment). The classification of the data would be at the four-digit level of ISIC or the national equivalent.

More-frequent-than-annual surveys

256. In the publication of data based on more-frequent-than-annual surveys, timeliness is of much greater importance than it is for infrequent inquiries, but the fact that the amount of data is very limited makes copy preparation and duplication much less of a problem. Data based on a sample minimizes the disclosure problem but also tends to make the final review more difficult,

since the cause of processing errors is harder to trace. The text should include a description of the sample and a table of sampling variances for the guidance of data users. Since, typically, the press is under time restraint in the publication of current data, it is normal for the text of the publication to include a summary analysis of the data in such a form that it can be picked up in its totality or used with only minor modification.

SAMPLING TECHNIQUES IN SURVEYS OF DISTRIBUTIVE TRADES AND SERVICES

1. Sampling is often desirable when statistics are needed but a complete census of the universe is not necessary. Taking a sample instead of a complete census usually costs much less and imposes a much smaller burden on the respondents. Sampling error is introduced, but as a result of a smaller workload the work may be done more carefully, possibly reducing the total error. Sampling may also make possible much earlier publication of the survey results.

2. Probability sampling means that every member of the universe has a non-zero probability of selection and that the final selections are determined by chance and not by anyone's judgement. If it is properly implemented, the sampling errors of a probability sample can be evaluated from the sample itself. Samples are often drawn in other ways, such as by taking those units that can be obtained most easily or those that, in someone's judgement, are representative of the universe. The sampling error of samples drawn in these ways cannot be computed; thus, their value is greatly reduced. The remainder of this discussion is, therefore, confined to probability sampling.

3. Since probability samples are the result of chance selection, a large number of results which vary in different degrees from the average (expected value) of these results is possible. This variation around the average result is called sampling error. The variance is defined as the average squared deviation of each result from the over-all result. The square root of the variance is called the standard error. Under the usually reasonable assumption of normality a distribution of errors can be inferred from the standard error. The standard

error divided by the estimate itself is called the coefficient of variation.

This measure of relative error permits the comparison of sampling error among estimates widely different in size and character. The square of this measure is called the rel-variance which is equivalent to the variance divided by the square of the estimate. The computation of sampling error described below is in terms of the variance which can be easily reduced to the other measures described, if desired.

4. Sampling error can be reduced by increasing the size of the sample (roughly speaking, the standard error varies inversely with the square root of the sample size). An approach often used in speculating about sampling error for samples of different sizes is to compute (or assume from past experience) a rel-variance for a single unit in the sample and then divide that value by the proposed number of units in the sample to obtain the rel-variance of the proposed design. For example, if we assume from past experience that the rel-variance per commercial establishment in a list sample is 1, then the rel-variance of a sample consisting of 100 establishments would be .01, and the coefficient of variation would be .1 or 10 per cent. (Note that neither the number of units in the universe nor their average size is needed for this computation.)

5. There is another kind of survey error called non-sampling error which is neither the result of the chance selection process nor, in general, affected by the size of the sample. Such error, for example, could result in deficiencies in the frame being used, errors in reporting, the failure of all units to report, errors in processing etc. Error of this type is particularly dangerous because it is not readily subject to measurement, as is sampling error. It should be kept under reasonable control by adequate quality-control programmes, even if it is necessary to sacrifice sample size in order to do this. In deciding on

the balance between the effort spent in reducing sampling error and that spent in reducing non-sampling error, it should be remembered that the objective is to produce the lowest possible total error for the available budget, not the lowest possible sampling error.

6. It follows from this principle that a simple design that can be implemented as planned is preferable to a more sophisticated design that, although theoretically more efficient, is subject to serious non-sampling error in its implementation due to its complexities.

7. The budget of any survey (and the sample size resulting from it) should theoretically be set at the point where the sum of survey cost plus the expected loss from wrong decisions made as a result of errors in statistics is at a minimum. It is usually impossible to determine the budget even roughly, principally because of the difficulty of evaluating probable decision loss. Additionally, budgets are often not under the control of statisticians. However, thought given to this principle may prevent major errors, such as taking a complete census when a sample survey will serve or taking a sample too small.

8. The first step in drawing any sample is to decide on a frame from which the sample is to be selected. In the case of commercial establishments, the frame is either a list of all such establishments or a list of units with which the commercial establishments can be associated (such as a list of the enterprises that own them or a list of the areas in which they are located).

9. Frames for samples of commercial establishments can be put into two general categories:

- (a) Establishments or the enterprises that own them (a list of which is often derived from tax records or trade association records). Samples drawn from this type of frame are hereafter referred to as list samples;

- (b) Areas of the country. To avoid bias, all areas should be given a chance of selection. Samples drawn from this type of frame are hereafter referred to as area samples.

Combination samples are also used if part of the universe is represented by a list sample and part of the universe is represented by an area sample.

10. The list sample usually has an advantage in terms of sampling error over the area sample for the following reasons:

- (a) The area sample requires a personal canvass to bring the listing of commercial establishments in the selected areas up to date. Canvassing for list samples can often be done primarily by mail/telephone which is usually less costly;
- (b) List samples can usually be stratified more effectively than area samples;
- (c) Area sampling is cluster sampling which usually reduces the efficiency per case selected. List samples generally are not cluster samples.

11. While list samples usually have a distinct advantage over area samples from the standpoint of sampling error, they have serious problems with non-sampling error. It is often very difficult to give each commercial establishment one and only one chance of selection from a list-sample frame, for the following reasons:

- (a) The list-sample frame is never up to date for late births;
- (b) The list-sample frame usually omits by definition part of the universe of interest. Usually, but not always, these omissions are small establishments;

- (c) The list-sample frame may also omit some establishments that by definition should be included in it. For example, tax evaders will not be on tax records; non-members will not be on trade association records;
- (d) The list-sample frame often contains duplications; a single establishment may have two or more listings which may not be exactly identical in name, address or other identification;
- (e) In a list sample, establishments that fail to report may be imputed as operating when in fact they are out of business, or vice versa;
- (f) In a list sample, it is often the practice to remove from the mailing list those sampling units that are reported to be out-of-business. These reports may be erroneous or out of date; units that were out of business at the time of reporting may have been reactivated. This factor tends to produce a cumulative downward bias in estimates derived from list samples.

12. In general, the area-sample approach, since it is based on actual observation, provides a firm basis for making correct judgements as to whether commercial establishments are currently operating or not and for giving each of them one and only one chance of selection. However, the area-sample approach also has bias problems in coverage. The principal problems are as follows:

- (a) Some commercial establishments are non-recognizable as such from outside observation. In general, they will be missed by the enumerator;
- (b) The enumerator will make errors in the listing process. Generally, he will list fewer than he should. These errors can usually be minimized by a quality-control programme.

13. It is to be noted that coverage errors in list samples are more likely to be cumulative than coverage errors in area samples. Cumulative coverage errors are particularly troublesome, because they distort estimates of change as well as estimates of level.

14. The choice of a frame for a given survey should be made on the basis of which frame source will give the smallest total error for the available budget; this, in turn, is dependent on the conditions of the particular survey. However, the following guidelines may be helpful:

- (a) To a considerable extent, the choice is dependent on the quality of the available lists. If these contain serious omissions or duplications, area samples will probably be the better choice. Note in (d) below that the combination of a sample from an incomplete list and an area sample to cover the deficiencies of that list is not recommended;
- (b) Multiple-list frames with overlapping coverage should be avoided. Theoretically, the resulting probability of duplication can be allowed for, if it is known, for sample cases, but the practical problems of determining to what extent it exists are great;
- (c) If a simple good list or a group of non-overlapping lists that together form a good list is available, the choice is not so clear;
 - (i) For a survey conducted only once (for example, a sample in connexion with a census), the list sample is the obvious choice;
 - (ii) The list-sample approach may also be practical for infrequent periodic surveys in which a long period of time is allowed for tabulation (e.g., six months or more). This is true

because the survey provides time for solving coverage problems of the list-sample approach;

(iii) Coverage problems of the list-sample approach are at their worst for a frequent periodic survey (e.g., monthly or quarterly) which has only a short interval between the end of the survey period and tabulation (e.g., a month or less). Here the list-sample approach is almost certain to have a substantial coverage problem. Whether this is more important than the greater sampling error of the area-sample approach depends in part upon the size of the budget available. Since the non-sampling error of the list sample will not decline with larger sample size, it may be a large part of the total error for a high-budget survey but a much less important part of the total error for a low-budget survey;

(iv) Partly, the decision depends on whether the collection of reports by mail/telephone is feasible. If it is not feasible, personal visits to units of a widely spread list sample could be much more expensive than visits made for a clustered area sample;

(d) If an area sample is used, it is recommended that it be supplemented with a list of large establishments or companies taken with certainty. Since these certainty units are canvassed each month, it is easy to keep track of their status so that they can be properly "skipped" in the area sample. Additionally, the probability of duplication would require duplicate reporting by the respondent. Combining list and area samples when the list

is not taken with certainty is not recommended. Here the problem is deciding which units to skip among those in the area sample that have had a chance of selection. This is usually so difficult that a combination is not recommended.

15. The following paragraphs describe one procedure for drawing an area sample of commercial establishments. The first step is to decide on the nature of the ultimate cluster, defined as a group of m area-sample segments containing an average of \bar{N} commercial establishments each and included in a selected primary sampling unit (PSU). Some decisions that need to be made about the ultimate cluster are:

- (a) What size should \bar{N} be?
- (b) What size should m be?
- (c) What should be the size and nature of the primary sampling unit?

16. The answers to the above questions depend on the cost and variance conditions of the particular survey being considered. However, the following rules of thumb should lead to results that are reasonably close to the optimum in most situations:

- (a) It is recommended that \bar{N} , the number of in-scope commercial establishments in each segment, be as close to 6 as possible (4 if retail establishments only are considered). This segment size was found to be optimum by empirical investigation under conditions in the United States of America and may be used as a starting point in other countries. However, since cost and variance conditions may differ significantly in other countries, it is recommended that empirical investigations be undertaken to determine which segment size provides the greatest reliability per unit of expenditure;

- (b) It is recommended that the number of such segments selected for the sample in each non-certainty PSU be set as nearly as possible at m , the number of segments that can be easily handled by a single enumerator;
- (c) It is recommended that the size and nature of the primary sampling units be determined by the following considerations:
- (i) A measure of size, such as receipts of commercial establishments or population estimates from a past census, should be available for each of the PSU;
 - (ii) The PSU should not be so large as to require an enumerator to spend most of his time travelling from home to segment or between segments. (Perhaps the average home-to-segment trip should not exceed three quarters of an hour, if possible);
 - (iii) The PSU should not be too small or there may be considerable loss due to high intra-class correlation;
 - (iv) Within the above limitations, the PSU should be as heterogeneous as possible (for example, each should be a mixture of urban and rural areas, if possible).

17. The PSUs will most likely be political subdivisions or combinations thereof. If a household area sample is canvassed in the country, very possibly the same set of sample PSUs can be used for an area sample of commercial establishments. It is not recommended, however, that the sample of area-sample segments used for the household survey be used for the commercial-establishment sample, since the correlation between households and commercial establishments is very low at this level.

18. The next step is to determine ℓ , the number of ultimate clusters that will be drawn into the sample. This, of course, is determined by the budget available

for this purpose divided by the estimated cost per ultimate cluster. The size of the budget should be determined by considerations discussed in paragraph 7 above. The next procedure should be to speculate on the probable sampling error (coefficient of variation, or CV) for area samples of various proposed sizes. Compute \sqrt{ml} , the square of the number of segments of the proposed design. Divide 1 by this number to obtain an approximation of the CV of level, 0.2 by this number to obtain an estimated CV of month-to-month change and 0.4 by this number to obtain an approximation of the CV of year-to-year change. For example, suppose we are considering a survey consisting of 40 ultimate clusters (ℓ) with an average of 10 segments (m) each. This would provide an estimated 400 segments (ml). The CV of level would then be 5 per cent ($1 \div 20$), the CV of month-to-month change would be 1 per cent ($0.2 \div 20$) and the CV of year-to-year change would be 2 per cent ($0.4 \div 20$). These crude approximations can be refined when data from pre-tests or actual surveys become available. Once ℓ has been determined, the over-all sampling rate, P, can be computed.

$P = 6\ell m/N$, where N is the estimated total number of commercial establishments in the country. Once P is decided upon and samples are drawn at that rate, it should not be changed even though N is proved to be in error.

19. Identify all certainty PSU in the country. Certainty PSU will be those in which $S_j \geq \frac{S}{\ell}$, where S_j is the measure of size, S is the same measure of size for the country and ℓ is the number of ultimate clusters in the sample, as defined above. For example, suppose population is used as a measure of size in drawing PSU. If the estimated total population (S) is 10,000,000 and 100 ultimate clusters (ℓ) are to be drawn, then any PSU with a population exceeding 100,000 ($\frac{S}{\ell}$) should be taken with certainty.

20. Stratify all remaining non-certainty PSU. The strata should be as nearly as possible of the size $2S/\ell$, in terms of the measure of size (S) being used. In our example, we would strive for strata with populations of 200,000. PSU placed in a given stratum should be as nearly alike as possible.

21. Subdivide each stratum formed as indicated in the previous paragraph into two substrata. Again, these substrata should be as nearly equal in size as possible (now the target size in the example is S/ℓ , or 100,000).

22. Select one PSU from each substratum formed with a probability of selection proportionate to S. To continue our example, suppose we have a given substratum that includes PSU A with a measure of size of 10,000 (people, in this case), PSU B with a measure of size of 30,000 and PSU C with a measure of size of 60,000. We would arrange a cumulative sum as follows:

<u>PSU</u>	<u>Measure of size</u>	<u>Cumulative measure of size</u>
A	10 000	10 000
B	30 000	40 000
C	60 000	100 000

We would then select a random number between 1 and 100,000. Any number between 1 and 10,000 would select PSU A; any number between 10,001 and 40,000 would select PSU B, and any number between 40,001 and 100,000 would select PSU C. The probabilities of selection are 10,000/100,000 (or 0.1), 30,000/100,000 (or 0.3) and 60,000/100,000 (or 0.6) for PSU A, B and C, respectively.

Note: The above method of selection results in one sample PSU per stratum. As a result, it is impossible to obtain an unbiased estimate of the variance. An over-estimate of the variance is obtained by the so-called collapsed stratum method of variance computation described below. It assumes that the two PSU were drawn from the same stratum.

23. Schemes are available for selecting two PSU per large stratum with probability proportionate to size without replacement. However, these schemes are more complicated and have a higher actual variance than that calculated in the sampling plan described here.

24. Another possible alternative is to draw two PSU per large stratum with probability proportionate to size with replacement. This method provides an unbiased estimate of the variance, but the actual variance is higher than that calculated by the method described above.

25. After the certainty PSU have been identified and the non-certainty PSU have been selected, the next step in the area-sample process is to subdivide the entire area of the selected PSU into area-sample segments. Each segment should have the following characteristics:

- (a) Reasonably permanent boundaries (streets, roads, railroads, streams etc.) which are recognizable to enumerators;
- (b) Six commercial establishments, as nearly as possible. If the smallest possible segment with recognizable boundaries contains 10 or more establishments, it should be designated "non-take-all" with a measure of size equal to the estimated number of establishments divided by 6 (rounded to the nearest whole).

26. In order to perform the segmentation process, it is necessary to have maps that show recognizable features, such as streams, roads etc., and the location of commercial establishments. If such maps are already available (for example, from a past census of commercial establishments), they should be used. It is not necessary that they be completely accurate or up to date. If such maps do not exist, they should be created by a personal canvass of sample PSU. Starting with maps that are as detailed as possible (enlarged, if necessary), the enumerators

should list all commercial establishments (as determined from outside observation) and add any recognizable features not already shown on the maps.

27. With maps already available or created as indicated in the previous paragraph, the sample PSU should be segmented in the office according to the rules in paragraph 25. The entire area of the sample PSU should be accounted for in this segmentation process.

28. Segments in sample PSU (certainty and non-certainty) should be numbered systematically (for example, in a serpentine fashion). In numbering, each measure of a non-take-all segment should be treated as a separate segment.

29. The within sampling rates should now be computed. For certainty PSU the within sampling rate is P , the over-all sampling rate. For non-certainty PSU the within rate is $P_{ghj} = P \cdot S_{gh}/S_{ghj}$.

30. In the above equation, S is the measure of size used in the PSU sampling process, S_{gh} is the measure of the h substratum within the g -th (large) stratum and S_{ghj} is the measure of the j -th PSU within that stratum. S_{gh}/S_{ghj} , therefore, is the inverse of the probability of selecting the PSU.

31. To continue our example, let us suppose that PSU A was selected; in that case, $S_{gh}/S_{ghj} = 1/0.1$, or 10. If we assume an over-all sampling rate (P) of 1 in 100, the within sampling in PSU would be $(1/100) 10$, or 1 in 10. The sample is actually selected by using the inverse of the within probability of selection, $1/P$ or $1/P_{ghj}$, as the interval, with a random starting point between 0 and this number. To select the segments, the resulting pattern is applied to the systematically numbered segments referred to in paragraph 28. Segments in each certainty PSU are systematically assigned as drawn to five random groups to facilitate the computation of variances described in paragraph 63.

32. It is usually worth while to supplement the area sample with a list of very large commercial establishments (or enterprises). This list can be obtained from any source, but all members of it should be selected with certainty and enumerated for each survey period. The cut-off for this list can be initially set at $S/6n\ell$, or 5 divided by the estimated total number of establishments in the area sample. S is some measure of size (sales, employment, payroll) available at both the national level and for each large establishment. It need not be the same measure used in sampling PSU. For example, if the estimated total retail receipts in a country is \$1 billion and we expect 2,000 area sample cases, then the certainty cut-off would be $\$1,000,000,000/2,000$ or \$500,000, and we would strive to include in our certainty list all establishments of the larger size.

33. This is a very crude approximation which essentially assumes that the larger variance and larger cost per case of area-sample establishments (as contrasted to list-sample establishments) cancel each other out in the optimum allocation formula. A more exact cut-off can be determined empirically by experimenting with different cut-off levels after some area sample listings are available. In these experiments, the total cost should be equalized by making the area sample larger or smaller, as necessary to compensate for the varying size of the list sample.

34. The instruction to the enumerator should be to list all commercial establishments found within the boundaries of sample segments, with the following exceptions:

- (a) Establishments in the certainty list should be skipped in the area sample. To facilitate this process, the enumerator should be given a list of certainty establishments (or enterprises) which are known to operate in his PSU or might do so.

(b) In non-take-all establishments, listed establishments should be sampled at the rate of k in K , where K is the measure of size assigned the segment and k is the number of sample "click-offs" (usually one) falling in the segment. The sampling can be done in the field by the enumerator or in the office. If done in the field, the starting point and interval should be predesignated and the enumerator should be given rules for order of listing that leave no room for judgement. Office sampling entails less risk of bias but may involve significant time and cost losses.

35. A subsample of each enumerator's listing and enumeration should be checked for quality. This check is more effective if it is done independently, i.e., the enumerator and rechecker prepare independent listings and reconcile them. The rechecking is also more effective if it is not done with a fixed pattern that permits the enumerator to know which months will be checked.

Note: The topics of weighting, variance computation, use of ancillary information in estimation procedures and rotation of samples apply to both list and area samples and are discussed in paragraphs 58 through 95.

36. If it is decided to use the list-sample frame, the first step is to stratify the list. Stratification means placing the members of the list into classes (strata) from which independent samples will be drawn. The objective should be to make sampling units within strata as nearly alike as possible; likely criteria for stratification are major kind of business and size, if available. Five or six non-certainty size strata within each kind of business should provide most of the gains, without being too complex. However, if the stratification is to be done by an enumerator in the course of the canvass, a single non-certainty stratum is recommended for simplicity. The objective

should be to make the non-certainty strata as nearly equal in size as possible. However, for the sake of simplicity it might be decided to use the same size boundaries for each kind of business. This would make the size strata equal at the level of all kinds of business but unequal within individual kinds of business.

37. The total size of a list sample will be determined by the total available budget (excluding fixed costs) divided by the cost per unit in the sample. The total size of the budget should be determined by considerations discussed in paragraph 7. In speculating about sampling error, it can be initially assumed that the coefficient of variation of a level estimate from a list sample will be 1 divided by the square root of the number of units in the sample. The coefficient of variation of month-to-month change might be in the neighbourhood of 0.3 divided by the square root of the number of units in the sample; 0.5 divided by the square root of the number of units in the sample would provide a rough approximation of the coefficient of variation of year-to-year change. (See paragraph 18 for an example of how to make these computations but use the number of units to be drawn in the sample from the lists). If the certainty class is important, make these approximations for the non-certainty class and then multiply the results by P_{NC} , the proportion of total receipts accounted for by the non-certainty class. These crude initial approximations can be refined as data from the survey become available.

38. A simple method for allocating the total sample among the various kinds of business is to make the allocation proportionate to the total measure of size of each kind of business. This will produce a near-optimum over-all total. The sampling error (coefficient of variation) for each kind-of-business estimate will be roughly inversely proportionate to the square root of its size. Of course, this distribution can be altered if it is felt that some kinds of business are

more or less important than this allocation would indicate.

39. The optimum allocation of the sample to a given kind of business among the various size strata is provided by the formula:

$$n_h = \frac{N_h \sigma_h}{\sum_{h=1}^R N_h \sigma_h} n_o$$

In the above formula, n_o is the total sample allocated to the kind of business (as per paragraph 38 above), n_h is the sample allocated to the h-th of R size strata within the kind of business, N_h is the number of sampling units in the h-th size stratum while σ_h is the standard deviation (for a single draw) of the h-th stratum. If the n_h indicated by this formula is equal to or greater than the N_h units in the stratum, all are taken with certainty.

40. A simple and often good approximation of the above allocation is the following:

- (a) Set the certainty cut-off at X/n_o , where X is the total measure of size for the kind of business and n_o is the sample allocated to the kind of business;
- (b) Allocate the remaining sample proportionate to the X_h (total measure of size in the h-th size group) in the given kind of business. For example, suppose a sample of 100 establishments has been allocated to the kind of business that has a total measure of size (e.g., payroll) of \$100,000,000. Then any establishment with a payroll of \$1,000,000 or more would be taken with certainty. If there were 10 such establishments accounting for \$20,000,000 in payroll, then there would remain a sample of 90 non-certainty establishments to be distributed.

If there were five non-certainty strata, each would be allocated a sample of 18 establishments.

41. The above approximation assumes that the coefficients of variation (for a single draw) are the same for all strata. Actually, the coefficients of variation are usually larger for the smaller-size classes. A somewhat better allocation is sometimes obtained by adding a small increment to each measure of size before proceeding as described in paragraphs 39-40 above. For example, if employment is being used as the measure of size, one employee might be added to the measure of size for each establishment (or enterprise, if this is used as the sampling unit) before the sample is allocated.

42. To draw the indicated sample in each kind-of-business size stratum, first list and number systematically all sampling units in the stratum (by size, kind of business, location etc.). Then select a sample by applying the pattern obtained by using the interval $I_h = \frac{N_h}{n_h}$ (rounded to the nearest whole number) and a random start between 1 and I_h . The sampling units should be assigned systematically, as drawn, to the five random groups or replications for purposes of the variance computations to be described later.

43. As indicated, the principal problems of sampling a list for a recurring survey are to ensure that each establishment has one and only one chance of selection and to make correct decisions as to which members of the sample are currently operating. In the following paragraphs, recommendations are made to minimize these problems.

44. A universe register of all units ever given a chance of selection should be maintained. Births and other suspected new units can be checked against this list to see if they have had a prior chance of selection.

45. A reliable source of births should be available and these should be sampled and added to the sample as quickly as possible.

(a) If reliable information for stratification (kind of business and/or size) is not available at time of birth, two-stage sampling should be used. A large initial sample should be drawn to obtain stratification information. Then a smaller sample should be drawn from the first sample so that the product of first and second probabilities is appropriate for the final stratum. The size of the first-stage sample of births (i.e., the rates at which that sample is drawn) is not subject to explicit mathematical optimization. As with other optimization problems of this type, one can experiment empirically as soon as data from the survey become available. This experiment should proceed along the following lines:

- (i) From the sample cases, recreate a universe by weighting each sample case by the inverse of the probability of selection. This universe would include both the original sample and births;
- (ii) Experiment with various first-stage sample rates for the births in the simulated universe created above. Note that the higher first-stage sample rates will result in fewer outliers (caused by cases whose first-stage probability is smaller than that called for by its final size and kind-of-business stratum). However, the higher first-stage rates will result in a greater cost due to the larger sample;
- (iii) At each simulated first-stage sample level, estimate the cost of the first-stage sample. Adjust the total sample size to allow for these differences in first-stage sample costs, so

that the total cost is held constant;

- (iv) Compute the over-all variance at each level using the indicated probabilities and select the design that provides the optimum balance between the effects of outlier births and total sample size.

Until survey data are available, the initial probabilities of the first-stage sample of births should be high enough to make it unlikely that outliers will result. For example, if it is possible that births will be large enough to justify selection with certainty, then the initial first-stage sample should be taken with certainty. Having a large initial first-stage sample of births will also make the experimentation described above more effective, since the simulated universe of births will be near the true universe of births;

- (b) Two main systems for birth sampling are used. In one, only true births (i.e., births with no predecessors) are sampled; successors are represented by taking the successors of existing sample members. In the other system, all births are sampled and all existing sample members are dropped when they go out of business, whether they have a successor or not. Since there is no clear-cut definition of successors and it is not always easy to get reliable information on them, the system of sampling all births is probably less subject to cumulative bias.

46. It is recommended that an investigation of the status of non-reporters in the survey be carried out to determine whether they are still operating. If this is done, the correct decision can be made as to whether to impute a non-zero

estimate for the unit. It is also recommended that deaths be surveyed at least once a year to determine whether they have been reactivated.

47. It is recommended that the sample be frequently redrawn for the following reasons:

- (a) To reduce the effect of cumulative biases;
- (b) To spread the burden of reporting;
- (c) To avoid variance increases due to changes in measures of size.

The sample can be drawn independently or with maximum or minimum overlap with the previous sample. Drawing an independent sample is by far the simplest way and achieves most of the rotation possible.

48. Redrawing a list sample is likely to result in a discontinuity between old and new results. This is one of the disadvantages of the use of list samples for recurring surveys. Of course, the area sample also has to be redrawn eventually but that step is needed less often than it is for list samples and, because it is less subject to cumulative biases, is less likely to result in major discontinuities.

49. Drawing a sample from a complete census represents another application of list-sample principles to commercial establishments. It should be done (to save time and money for both respondents and the Government) when some of the data are not required on a small-area basis (e.g., receipts by commodity line, inventories, capital expenditures). This application avoids many of the problems described in paragraphs 43 to 48 above, because the complete census of basic data for items such as number of establishments and total receipts provides a one-time universe list which is relatively up to date, complete and unduplicated. However, this application has some special problems of its own which depend on the method of canvass used for the complete census and are described below.

50. There are three general methods for taking a complete census of commercial establishments. The first is by personal canvass in which enumerators create the list of census establishments in the course of personal enumeration. The second is to canvass (usually by mail/telephone procedures) a pre-existing list of commercial establishments often compiled from administrative records. A third method is to compile the desired basic data from administrative records themselves. The census may be taken by a combination of these three, using different methods for different parts of the universe. The sampling problems arising from each of these three complete census procedures are described below.

51. If a census (or part of a census) is taken by personal canvass with no pre-existing list, then the sample of establishments should be drawn during or after the canvass. Theoretically, the sample could be drawn before the census by having a sample of areas designated, but there seems to be little reason to accept the inefficiencies of the area sample when an up-to-date list will be available as a result of the canvass. If the sample is drawn in the course of the enumeration, it should be extremely simple and it should leave no leeway for the enumerator to exercise his judgement. Usually, it is desirable to have two strata, a certainty stratum and a single non-certainty stratum. For example, the rule might be to take in the sample with certainty all establishments with 10 or more persons engaged and a sample of 1 in 10 of those establishments with fewer than 10 engaged (the certainty cut-off and the rate for the non-certainty stratum would be determined by principles discussed in paragraphs 36 to 42). The enumerator should be given inflexible rules for determining which non-certainty establishments fall in the sample. For example, if the sample is pre-designated on the listing sheets, he should be given a route of canvass which should be followed without deviation. Alternatively, each questionnaire could have a sample designation under a seal.

The rule in this case would be not to lift the seal until the census of basic items is completed. When the enumerator removes the tape, he then determines if the establishment falls in the sample or not.

52. In spite of the above precautions, samples drawn by the enumerator in the course of the enumeration are almost certain to be biased, although often the bias is small and does not invalidate the results. The effect of enumerator bias (or bias caused by other factors) can often be reduced by the estimation procedure described in paragraph 57 below.

53. The alternative to having the enumerator draw the sample in the course of canvass is to have it drawn after the completion of the basic census. In this procedure, the enumerator mails the basic census questionnaires to the national or regional offices, where they are sampled (probably as received). This method makes possible the use of more complex and more efficient stratification schemes. It also reduces the risk of bias. However, it may be impractical for two reasons:

(a) It will increase the total amount of time required to take the census, since it introduces a delay between the basic census and the sample census canvassed. In addition to causing delay in publication, this factor may introduce bias, because a significant number of establishments may go out of business in the interval.

(b) It will increase the total cost, because a separate canvass effort (whether personal, mail or telephone) must be made to complete the sample census. If the enumerator does the sampling in the course of enumerating the basic census, he can pick up all or part of the sample questionnaires or at least leave them with the respondent.

54. If a census (or part of a census) is obtained by a canvass (probably mail/telephone) of a pre-existing list, then we have the choice of drawing the sample

before the basic census is mailed out or after the questionnaires with the basic information have been received in the office. In either case, the sampling is done in the office so that more complicated and efficient designs employing the principles stated in paragraphs 36 to 42 can be used. However, if the sampling is done before the basic census questionnaires are sent out, good information on size and kind of business needed for stratification purposes may be unavailable. This provides a particularly difficult problem if the sample questionnaire to be sent out depends on the kind of business of the respondent.

55. If sampling is done after the basic census questionnaires have been received in the office, better information on size and kind of business to use in the stratification process will be available. Again, however, this approach has severe timing and cost disadvantages owing to the fact that the canvass for the sample census must follow (perhaps by a significant lapse of time) the canvass for the basic census.

56. If all or part of the basic census is obtained by the compilation of data from administrative records, then again the choice is between drawing the sample before or after the data to be used in the census are received. Often an up-to-date list from which to draw a sample is not available until after the current data have been received. Often this is so late that drawing and canvassing a sample from the list becomes impractical. Therefore, sample data may not be collected for this portion of the universe, especially if it is a small part of the universe.

57. Regardless of how the sample is drawn and canvassed, the basic data (for example, number of establishments and total receipts) will be available for both the universe and the sample. Those data can be used to improve the sample estimates by use of a ratio estimate of the form $\frac{Y'}{X}$. In the above form Y'

is an estimate of some item available for the sample questionnaire only; the symbol X represents a correlated statistic obtained from the basic census; X' represents the same total as estimated from the sample. (The X chosen should be that believed to be most closely correlated with the Y estimate.) For example, suppose that Y' represents estimates of a particular commodity line (e.g., clothing in department stores) and X represents total sales in department stores. Then, a usually better estimate (i.e., Y'') of sales of clothing in department stores can be obtained from the estimate $Y'' = \frac{Y'}{X} X$. The use of this type of estimate has three advantages:

- (a) The Y'' (ratio) estimate may have a lower variance than Y' (simple estimate) if there is sufficient correlation between X and Y ;
- (b) The estimate can be used as a means of imputing non-responses for the sample (Y) data. This assumes that the X (census total) is complete;
- (c) The procedure can be used to alleviate the effect of biases in the sampling process, introduced by enumerators or from other sources. This method assumes that the ratio of Y to X is the same for the true population as for the imperfectly drawn sample. While this is not necessarily true, it is often nearer the truth than the alternative of assuming that the simple Y' is unbiased;
- (d) If total receipts are broken down into components (such as commodity-line sales), then the sum of the components will add to the total receipts in the basic census (provided that X' is made to equal Y' where the sum is taken over all commodity components).

In connexion with advantages (b), (c) and (d) above, this procedure should not be used to hide significant defects of a sample census. If the X' (simple) values vary from the X (census) values more than could be reasonably expected from sampling error, then users should be informed of the ratios $\frac{X'}{X}$ used, so that they know how much correction of sampling biases and/or imputation has been accomplished by this procedure.

58. The following paragraphs discuss some principles of sampling which apply to both list samples and area samples. The first topic concerns the weighting process. Other topics discussed, in order, are computation of the variances, the rotation of samples, the use of ancillary information to reduce variances and the perils of arbitrary weight reduction.

59. Weighting to produce simple unbiased estimates is always by the inverse of the probabilities of selection. In the case of the area samples, the probability of selection is the product of the probability at the various stages. In the case of a two-stage sample, this would be the probability of drawing the primary sampling unit times the probability of drawing the segments within the primary sampling unit.

60. In the example given in paragraphs 29 to 31, the PSU was selected with a probability of 1 in 10 and the segments within the PSU were drawn at a probability of 1 in 10. It is no coincidence that the over-all probability of 1 in 100 is equal to the pre-designated rate (P) of 1 in 100. The weight to be used in this case is 100. If the area sample is drawn as recommended, then the probability of selection (P) is uniform for all parts of the sample, and the results are weighted by $1/P$. (A sample where the weights are uniform is often called a self-weighted sample.)

61. In the case of the list sample, the weight is the interval (I_h) used in the selection process (see para. 43) or 1, if taken with certainty.

62. The simple unbiased (weighted) estimates are always used either as the final estimates themselves or as the building blocks of the more complicated estimates described below. They are also used in the variance computations, again either directly or as building blocks, if one of the more complicated estimation procedures is used.

63. For designs recommended here (both area sample and list sample), the variance of the simple unbiased estimate can be computed from the following formula which is appropriate for any stratified sample:

$$\sum_{h=1}^R (1-f_h) \frac{n_h}{m_h-1} \sum_{j=1}^{n_h} (Y'_{hj} - \bar{Y}'_h)^2 \quad (1)$$

In the above formula, h is the h -th stratum (of R total strata). In the area sample each of the certainty PSU constitutes a separate stratum. In the non-certainty portion of the area sample, each of the original strata from which two ultimate cluster draws were taken constitutes a stratum, regardless of whether these two ultimate clusters were drawn by creating two substrata and drawing one from each or by drawing the two ultimate clusters from the original stratum (with or without replacement).

In the list sample, each kind-of-business size cell from which an independent sample was drawn constitutes a stratum, and f_h is the sampling fraction used, the probability (or inverse of weight) used in the h -th stratum. Where the collapsed stratum technique is used (i.e., where only one draw per substratum has been made), f_h is set at zero. Y'_{hj} is the simple weighted value of the j -th replication in the h -th stratum. In one-stage sampling, Y'_{hj} is the weighted value of the j -th group of sample draws which were created in the sampling process. In multi-stage sampling, Y'_{hj} is the weighted value of the

j-th ultimate cluster in the h-th stratum whether the draws were actually made from the strata or (as in the collapsed stratum method) by drawing one ultimate cluster from each of two substrata.

The number of the group totals described above in the h-th stratum is m_h . In the non-certainty portion of the area sample, $m_h = 2$ if drawn as recommended.

\bar{Y}'_h is the simple average of the Y'_{hj} canvass described above, that is:

$$\bar{Y}'_h = \frac{\sum_{j=1}^{m_h} Y'_{hj}}{m_h}$$

If the estimate used is more complicated than the simple unbiased estimate, then the variance is calculated from the same form as above but with U'_{hj} substituted for the Y'_{hj} described above. The U'_{hj} are linear combinations of the Y'_{hj} . The U'_{hj} for the particular complicated estimates referred to below are given. For a general method of computing the U'_{hj} for any complicated estimate, the reader is referred to the bibliography; the method has been computerized in a program available from the Bureau of the Census of the United States of America. This program computes variances of complicated estimates with no more input than that required to obtain the estimate itself.

64. Samples for recurring surveys of commercial establishments are often rotated over time. There are two reasons for rotating such samples:

- (a) To spread the burden of reporting among more respondents;
- (b) To reduce the sampling error of the survey estimates.

65. The recommended periodic renewal of list samples to reduce cumulative bias (paras. 47 to 48) may accomplish the purpose of spreading the burden of reporting without adding to the rotation of the sample. Since cumulative bias in an area sample is a less severe problem, the sample renewals may be too infrequent to

accomplish the desired spreading of the burden of reporting. Some additional rotation of the sample may be introduced to accomplish this purpose.

66. Rotation of the sample to accomplish variance reduction is rather complicated and therefore carries the risk of additional bias. A non-rotating panel is recommended if there is any doubt about the ability to handle these complications properly. Following are the principles of a basic rotation plan. For more detail, see the bibliography.

67. Using a limited number of rotating panels (called a closed set) over and over again is recommended. This has two advantages over using a new panel for each new reporting period:

- (a) It costs a considerable amount of money to prepare a panel for enumeration, whether it is a list panel or an area-sample panel. By limiting the number of panels, we limit the initial cost;
- (b) If we use a closed set of panels, we can apply the large observation procedure described below which can be of considerable value in reducing variance.

68. A convenient number of panels to be used in a closed set for a monthly survey is 12. This number is convenient because one panel can be designated the January panel to be enumerated in January of every year, one the February panel and so on through December. This means that the estimates for any month of the year will be based on the same set of panels for all years.

69. This will improve the important estimates of year-to-year change because of the year-to-year correlation which exists. If the initial costs are too high for 12 panels, then a smaller number can be used. If the number of panels is 2, 3, 4 or 6, the same panels will still be repeated each year; if 3 are used, the same set of panels will be repeated each quarter. Of course, reducing the number of panels in the set reduces somewhat the gains that can be made, but

even a set of as few as 4 panels would extract most of the variance gain available from rotation.

70. There are two different types of patterns of reporting that can be used with a set of rotating panels. If two months of data can be obtained from each reporter at each canvass, then the entire panel should be replaced the next reporting period. To illustrate this, we will talk about the January panel of a set of 12 monthly panels. In early February when the canvass would take place, if we can get the respondent to supply both January and December information (for example, receipts for these two months), then we would drop the January panel completely for the next canvass and canvass only the February panel. If, on the other hand, we could not obtain reliable information about December sales in early February, then we would request only the January sales from the January panel. However, in order to obtain sales from identical establishments, we would retain the panel for one more month and ask for their February sales at the time of the canvass of the February panel early in March. With this plan, we are canvassing each panel twice a year instead of once a year. If we wanted to equalize canvass effort with the first plan, then each panel could be only half as large.

71. Obviously, the first plan provides more information per unit in the sample. Using it, we can base our measure of month-to-month change for identical establishments on the entire panel instead of on half of it. However, the first plan involves an additional bias hazard which is not trivial. The report obtained in the current month is required to have the same expected value as the report obtained in the previous month. We again use the month of January as an example. Two reports relating to January activity will be obtained. One (the current report) will be obtained from the January panel immediately after the end of that month. Another estimate for the month of January will be obtained from the February panel

early in March. There are a number of reasons why the expected value of these reports may differ. First, if the reports come from records, the records for January may be more complete by early March than they were in early February, resulting in different reported levels of sales. Secondly, if the owners of the establishment estimate the reports, they may well tend to report January sales in each March according to business conditions at that time rather than according to business conditions in January. And thirdly, it is very difficult to treat births and deaths properly when two months are obtained in one canvass. For example, suppose a business is operating in early February but is out of business by early March. In an area-sample enumeration, the enumerator would list the establishment and record its sales for January as part of the canvass for the January panel. However, if this establishment fell in the February panel, its January sales would probably be missed, because the establishment would not be in existence at the time of the canvass of that panel.

72. In spite of the attracting gains of the first type of reporting pattern, it should not be used until empirical evidence has been acquired that the reports for the current and previous month have essentially the same expected value. This is especially true because the composite estimate described below magnifies whatever difference exists in the simple unbiased estimates.

73. One method of reducing the variance of the level estimates with rotating panels is to use the composite estimate, by which we use the information that has been reported in the previous periods to strengthen the current estimate. The form of the composite estimate depends on the type of reporting pattern. For the first type, the composite estimate which we will call Y_1''' (the subscript 1 refers to the data month being estimated) has the form:

$$Y_1''' = (1-K)Y_1' + K \frac{Y_1'}{Y_{1-1}''} Y_{1-1}''' \quad (2)$$

in the above estimating form we will again assume that we are obtaining an estimate for the month of January. In this case:

Y_1''' = the composite estimate for month 1 (January)

Y_1' = the single unbiased current estimate for the i-th month (the simple unbiased estimate for January from the January panel)

Y_{1-1}'' = the simple unbiased previous month's estimate for month i-1 (the simple estimate for the month of December as derived from the January panel). Note that the estimates Y_1' and Y_{1-1}'' are therefore derived from the same panel

Y_{1-1}''' = the composite estimate for the previous month (December) which was derived from the same formula

K = a constant which determines the relative weight placed on the two estimates being combined. It, in turn, depends on the month-to-month correlation of the simple estimates derived from the same panel. If we call this correlation

ρ the optimum K is
$$\frac{1 - \sqrt{1 - \rho^2}}{\rho}$$

If K is set at 0.72, it will be optimum when $\rho = 0.95$ and will be near optimum when ρ is near this value. When $\rho = 0.95$, the variance of the composite estimate (Y_1''') will be about one third the variance of an estimate from a non-rotating sample of the same size.

74. If the second type of reporting pattern is used, we will have two simple unbiased estimates available for the month of January (one from the January panel which we will call Y_1'' and one brought forward from the December panel which we will call Y_1'). Note, however, that each of these estimates is based on half-size panels. The composite estimate in this case takes on the following form:

$$Y_1''' = K \frac{Y_1'}{v_{1-1}} Y_{1-1}''' + LY_1' + (1-K-L)Y_1'' \quad (3)$$

Note again the Y_1' and Y_{1-1}'' are from the same (December) panel while Y_1'' is from the January panel. Y_1''' is the composite estimate as it was computed from a formula of this type for the month of December.

75. The optimum constants (again in terms of month-to-month correlations of identical panels) are:

$$K = \frac{1 - \sqrt{1 - \rho^2}}{\rho}$$

$$L = \frac{(1 - \sqrt{1 - \rho^2})(1 - \rho)}{\rho^2}$$

If the month-to-month correlation is at the level of 0.95, these constants have the following values:

$$K = 0.72$$

$$L = 0.04$$

76. The composite estimate of level will have a variance a little less than half that of a non-rotating sample of the same size.

77. While the variances of the level estimates are substantially improved by a composite estimate of this type, the variances of the estimates of month-to-month change worsen. For the first type of reporting pattern with the assumed correlati

and constants, the variance of the estimate of month-to-month change is increased by approximately one third. The variance of this estimate is more than doubled when it is necessary to use the second type of reporting pattern. It should be pointed out that while the composite estimate causes a loss in the estimates of month-to-month change, these may be compensated for by the often very substantial reductions that can be achieved through the use of the large observation procedure described in paragraphs 84 to 87.

78. Obviously the use of rotating panels and the composite estimate procedure are much more worthwhile if it is feasible to obtain two months of information with each canvass (i.e., first type of reporting). If this is not the case, it could well be that the complications of this system are not worthwhile.

79. Variances for the composite estimate (formulas 2 and 3) are computed according to the general rule in paragraph 63, that is, a substitute variable U'_{hj} is produced for each sample draw in each stratum and this value is substituted for the Y'_{hj} in the variance estimation formula (formula 1). The form the U'_{hj} takes for estimates from the first type of form of composite estimate is:

$$U'_{(1)hj} = Y'_{(1)hj} + (K) \frac{Y'_j}{Y_{1-1}} \left[U'_{(1-1)hj} - Y''_{(1-1)hj} \right] \quad (4)$$

K is the generalized constant which we suggested setting at 0.72.

80. The symbols have the same meaning as in formula 2 except that the subscript hj means that only weighted data from the j -th draw of the h -th stratum are included. Note, however, that the ratio $\frac{Y'_1}{Y_{1-1}}$ is still computed from the over-all totals.

81. It can be shown that the sum

$$\sum_{h=1}^R \sum_{j=1}^{M_h} U'_{(i)hj}$$

is exactly equal to the Y_i''' computed in formula 2. Therefore, the $U'_{(i)hj}$'s are computed for each draw and are simply summed to obtain the desired national totals. It is also to be noted that this is an interactive process, i.e., the U'_{hj} 's computed for month i are used as part of the data input for computation of U'_{hj} 's for month $i + 1$.

82. For estimates derived from the second type of pattern (formula 3), the U'_{hj} 's take the following form:

$$U'_{(i)hj} = (K+L)Y'_{i(hj)} + K \frac{Y'_i}{Y_{i-1}} - (U'_{(i-1)hj} - Y''_{(i-1)hj}) + (1-K-L)Y''_{(i)hj} \quad (5)$$

83. The symbols have the same meaning as in formula 3 except that, again, the subscripts hj mean that only weighted data from the hj -th draw are used. The constants K and L have the values 0.04 and 0.72, respectively, if the rule-of-thumb approximations described in paragraphs 73 to 78 are used.

84. If a closed set of rotating panels is used, such as 12-month panels, a large observation procedure can be used to reduce the weights of outliers in an unbiased fashion. This procedure can produce some significant variance reductions, particularly for the month-to-month ratio. This is important because the variance of the month-to-month ratio is increased by the use of the composite estimation procedure, as indicated in paragraphs 73 to 78.

85. Again we assume 12 panels. The weight of any outlier can be reduced to one twelfth of its original size for any establishment that is above the cut-off. To accomplish this, of course, the establishments must be enumerated every month instead of once every 12 months. This will be an unbiased procedure, provided that the establishment will receive the same treatment regardless of which of the

12 panels it happens to fall into. To ensure the same treatment, we use the following procedure: Set a cut-off for each month of the year; for the i -th month, we will call it C_i . This cut-off can be the same for all months of the year, but if there is a definite seasonal pattern, it would be better to vary the cut-offs approximately according to this pattern. The cut-off should be in terms of weighted values (of receipts or some other characteristic). The optimum cut-off is best determined by empirical experimentation with actual data as soon as they are available (under fixed cost conditions) but, as an initial approximation, it is suggested that the cut-off be set at a level equal to five times the expected average weighted monthly sales from the sample.

86. Once the cut-offs have been determined, the weighted values for each establishment in a given monthly panel are compared with the cut-off for that particular month. If the weighted receipts are above the cut-off for the appropriate month, the establishment is placed on a "potential large observation" list. After this has been done for all 12 panels, the entire potential large observation panel is canvassed for the missing 11 months of receipts data. If the receipts are above the cut-off for all of the 12 months, the establishment qualifies for the "permanent large observation" panel and is surveyed for every panel, its weight being divided by 12.

87. If the large observation procedure is used in conjunction with the composite estimate procedure, no changes need be made in the form of the U_{hj} used for variance computation. The weighted values from this panel should be included in all the simple unbiased Y values shown (both Y' and Y'' values).

88. If an area sample is being used, it is unlikely that the non-certainty PSU will be rotated on a frequent basis. The rotation of the area sample is likely to involve only the rotation of sets of segments within PSU. In this

case, the same formulas and procedures are used but only the within PSU component of the variance will be affected.

89. Often ancillary information, such as employment, payroll and so on, reported in recent tax records or data on receipts or other statistics reported in a recent census, can be used to improve the current estimates. In order to be useful, these data must be available for the universe and it must also be possible to evaluate them for the sample that was drawn.

90. The recommended estimation form for utilizing this ancillary information is:

$$Y'_R = \sum_{h=1}^L \frac{Y'_h}{X'_h} X_h \quad (6)$$

where Y'_R is the adjusted estimate, Y'_h is the simple weighted current estimate, X'_h is the weighted sample estimate of the correlated variable (X_h being the universe total of this variable) and h is the h -th of L major kind-of-business classes at the national level. (Note that the h classes here may represent combinations of the R strata used in the selection process.)

91. The following observations are made about this estimate:

- (a) In general, if the correlation between X and Y exceeds 0.5, Y'_R will have a lower variance than Y' (the simple estimate);
- (b) The L kind-of-business classes for which ratio estimates are made should have reasonably stable estimates of X'_h (i.e., a coefficient of variation of 15 per cent or less). Classes that do not meet this requirement should be combined with other classes;

- (c) If the data represented by the symbol X were used in the stratification and sampling processes, there may be little additional gain from the ratio estimates;
- (d) If this form of ratio estimate is used, the form of the U'_{hj} used in the variance formula became $U'_{hj} = Y'_{hj} - R'_h X'_{hj}$ where R'_h is the appropriate ratio of Y'_h to X'_h in formula 6.
- (e) Where the area sample is used, data for the correlated statistics are usually not available at the segment level. In this case, the X'_{hj} 's will be weighted PSU totals and will affect only the between-stratum component of the variance;
- (f) If the composite estimate is used to reduce the within PSU variance of an area-sample design and the ratio estimate is used to reduce the between PSU variance, then the resulting estimate (which we will call Y'''_R) would take the following form:

$$Y'''_{(i)R} = \sum_{h=1}^L \frac{Y'_{(i)h}}{X'_h} X_h \quad (7)$$

$Y'_{(i)h}$ is the composite estimate described in paragraphs 73 to 78 for the h-th major kind of business, X'_h is the estimate for the h-th major kind of business for the ancillary statistics obtained from weighted PSU totals for the h-th kind of business and X_h is the universe total of the statistic for the h-th kind of business.

92. If the above estimate is used, then the U'_{hj} value used in the variance formula (we will call this U''_{hj}) will have the following value:

$$U''_{hj} = U'_{hj} - R'_h X'_{hj} \quad (8)$$

In the above formula, U'_{hj} has the same value it has in formula 4 or 5, depending on the type of rotation, R'_h is the appropriate $\frac{Y''_{(1)h}}{X'_h}$ ratio for the hj -th draw and X'_{hj} is the weighted PSU total for the hj -th ultimate cluster.

93. A note should be inserted warning against large-scale arbitrary reduction of the weights of outliers in any type of sample. This is a great temptation in a sample of establishments, because an establishment (or enterprise) can become very large between the time it is subjected to sampling in a list sample and the time it is enumerated. Also, it is possible for very large establishments to appear in the area sample, since establishments are not stratified (except for a possible certainty list). For example, suppose a large department store appeared in the sample for city A with a weight of 100. Since city A is a city of moderate size, it is known that there could not be 100 department stores in that city. Therefore, the temptation is to reduce arbitrarily the weight (probably to 1). This would probably be the correct action to take if the sole purpose of the survey were to estimate department stores in city A. However, the over-all national total is usually of much greater importance and in this universe there could easily be 100 or more outliers of this type. To eliminate this case or reduce its weight would produce a significant underestimate.

94. The fact is that there is no way to tell from the sample itself whether extreme cases are over-represented or under-represented in the particular sample. Usually, information available outside the sample is too limited to permit this

judgement at the over-all level. Therefore, arbitrary weight reductions should be made very sparingly, since only a few such reductions can produce a significant downward bias.

95. The large observation procedure described in paragraph 86 is a method for reducing the weights of outliers in an unbiased fashion and can be of considerable importance in surveys where outliers present a major problem.

Note: The discussion above is specifically directed and mainly limited to practical techniques for applying sampling in distributive-trade surveys. It would be valuable to read also the documents on sampling listed in the bibliography which give some other methods that may be useful and discuss some related issues that are not covered here.

Annex II

BIBLIOGRAPHY

International Recommendations on Statistics of the Distributive Trades and Services, Series M, No. 57. 55 pp. United Nations publication, Sales no. E.75.XVII.9.

International Standard Industrial Classification of All Economic Activities (ISIC), Series M, No. 4, Rev. 2. United Nations publication, Sales no. E.68.XVII.8.

Indexes to the International Standard Industrial Classification of All Economic Activities, Series M, No. 4, Rev. 2, Add. 1. 247 pp. United Nations publication, Sales no. E.71.XVII.8.

Standard International Trade Classification, Revision 2 (SITC), Series M, No. 34, Rev. 2. 117 pp. United Nations publication, Sales no. E.75.XVII.6.

Classification by Broad Economic Categories, Series M, No. 53, Rev. 1. 53 pp. United Nations publication, Sales no. E.76.XVII.7.

"International standard classification of all goods and services" (E/CN.3/493).
(To be issued as a United Nations publication)

Bibliography of Industrial and Distributive-trade Statistics, Series M, No. 36, Rev. 4. 177 pp. United Nations publication, Sales no. E.75.XVII.7.

Recommendations for the 1973 World Programme of Industrial Statistics, part III. Organization and Conduct of Industrial Censuses, Series M, No. 54. 111 pp. United Nations publication, Sales no. E.72.XVII.10.

A Short Manual on Sampling, vol. I. Elements of Sample Survey Theory, Series F, No. 9, Rev. 1. 325 pp. United Nations publication, Sales no. E.72.XVII.5.

A Short Manual on Sampling, vol. II. Computer Programmes for Sample Designs, Series F, No. 9. 78 pp. United Nations publication, Sales no. E.71.XVII.4.

A Short Manual on Sampling, vol. II. Computer Programmes for Sample Designs, Series F, No. 9, Add. 1. 46 pp. United Nations publication, Sales no. E.73.XVII.8.

Recommendations for the Preparation of Sample Survey Reports (provisional issue), Series C, No. 1, Rev. 2. 16 pp. United Nations publication, Sales no. E.64.XVII.7.

Sample Surveys of Current Interest (twelfth report), Series C, No. 13. 477 pp. United Nations publication, Sales no. E.73.XVII.5.

United Nations publications are available from the Sales Section, United Nations, New York or Geneva.

كيفية الحصول على منشورات الأمم المتحدة

يمكن الحصول على منشورات الأمم المتحدة من المكتبات ودور التوزيع في جميع أنحاء العالم. استعلم منها من المكتبة التي تتعامل معها أو اكتب إلى: الأمم المتحدة، قسم البيع في نيويورك أو في جنيف.

如何购取联合国出版物

联合国出版物在全世界各地的书店和经销商均有发售。请向书店询问或写信到纽约或日内瓦的联合国销售组。

HOW TO OBTAIN UNITED NATIONS PUBLICATIONS

United Nations publications may be obtained from bookstores and distributors throughout the world. Consult your bookstore or write to: United Nations, Sales Section, New York or Geneva.

COMMENT SE PROCURER LES PUBLICATIONS DES NATIONS UNIES

Les publications des Nations Unies sont en vente dans les librairies et les agences dépositaires du monde entier. Informez-vous auprès de votre libraire ou adressez-vous à : Nations Unies, Section des ventes, New York ou Genève.

КАК ПОЛУЧИТЬ ИЗДАНИЯ ОРГАНИЗАЦИИ ОБЪЕДИНЕННЫХ НАЦИЙ

Издания Организации Объединенных Наций можно купить в книжных магазинах и агентствах во всех районах мира. Наводите справки об изданиях в вашем книжном магазине или пишите по адресу: Организация Объединенных Наций, Секция по продаже изданий, Нью-Йорк или Женева.

COMO CONSEGUIR PUBLICACIONES DE LAS NACIONES UNIDAS

Las publicaciones de las Naciones Unidas están en venta en librerías y casas distribuidoras en todas partes del mundo. Consulte a su librero o diríjase a: Naciones Unidas, Sección de Ventas, Nueva York o Ginebra.
