

PROPIEDAD DE
LA BIBLIOTECA C.1



UNITED NATIONS

ECONOMIC
AND
SOCIAL COUNCIL



LIMITED

ST/ECLA/Conf.32/L.8
17 May 1968

ORIGINAL: ENGLISH

SEMINAR ON THE ORGANIZATION AND
CONDUCT OF POPULATION AND HOUSING
CENSUSES FOR LATIN AMERICA

Organized by the United Nations Economic
Commission for Latin America, Statistical
Office and Office of Technical Co-operation,
with the collaboration of the Inter-American
Statistical Institute, the Latin American
Demographic Centre and the United States
Bureau of the Census

Santiago, Chile 20-31 May 1968

CHECKING, EDITING AND CODING OF CENSUS QUESTIONNAIRES

Prepared by ECLA Secretariat



TABLE OF CONTENTS

	<u>Paragraphs</u>	<u>Page</u>
I. INTRODUCTION	1- 8	1
II. CHECKING	9-21	4
A. Checking the coverage	13-16	5
B. Checking on entry omissions	17-21	7
III. CODING	22-35	11
A. Types of codes	22-31	11
B. Forms of coding	32-33	16
C. Quality control	34-35	19
IV. EDITING	36-43	21
A. Editing of entry omissions	38-39	21
B. Editing of inconsistencies	40-43	23
Annex I		25
Annex II		27

I. INTRODUCTION

1. After the census enumeration, begins a new stage of the census operation, i.e. the post-enumeration work. Because the information collected on the questionnaires constitutes the raw materials, which must subsequently be processed in order to obtain the desired result, i.e. the census tabulations, the first part of post enumeration work consists of the data processing operations.

2. The data processing operations can be divided in two principal parts:
 (a) preparation for the mechanical data processing;
 (b) mechanical data processing either by tradition or by electronic methods.

This document deals with the first part of this stage of the census operations; the mechanical (especially electronic) data processing methods are discussed in another document of this Seminar.^{1/} In this document "data processing operations" refers to the whole operation, including both parts (a) and (b).

3. There are various methods and forms by which individual countries organize their work in this stage, but generally three main operations could be distinguished:

- (a) checking of questionnaires;
- (b) editing of questionnaires;
- (c) coding of questionnaires.

In some cases these phases are prepared and organized separately, in other cases some combination is used between the individual phases.

4. The stages of processing and the order in which they are executed could differ somewhat according to the processing method used but all methods have in common the need to provide for editing of the original information for missing and inconsistent data, transcribing the information from the questionnaire for mechanical or electronic processing and tabulation of the results.

^{1/} See ST/ECLA/Conf.32/L.9

5. Fairly early during the preparation of the census, decisions should be taken on the location of the data processing centres to which the filled questionnaires are to be despatched from the field and where the preparation for data processing and also the mechanical data processing itself should be done.

The choice which has to be taken is between a centralized processing centre and a series of decentralized processing centres for this purpose. The centralized processing has the advantage of consistency of decisions and of simplicity of other controls. Decentralized data processing, especially in the preparatory phase, however, is almost imperative in very large countries where the number of census questionnaires and personnel required for processing could not be accommodated at a single site, i.e. in the central census office. Furthermore, decentralized processing can be done closer to the field, with the consequent benefits of easy accessibility for re-investigation. It is preferable, however, to establish an intermediate organizational form for data processing: the first-stage checking and eventually some part of the coding and editing could be done decentralized, but the more complicated part of coding and editing, especially the mechanical data processing work should be done centralized.

6. It is necessary to organize the data processing operations, in such a way that the first stages of post enumeration can be started immediately after the enumeration. This has to be underlined, because often in the period after enumeration there is a natural tendency for the staff to relax to a certain extent and to lessen its alertness. This is particularly so in countries where the census staff is small and enumerators might be used as coders in the office after the completion of enumeration.

To avoid such delay in data processing operations, it is essential to prepare in detail all materials and instructions before the enumeration, so that after completion it will not be necessary to wait for the preparation of these materials, but the work can start immediately.

7. A decision must be made also on the degree of division of labour to be employed. One extreme alternative would be that the first working phase, for instance coverage checking, would be performed for all questionnaires before starting with the next phase. The other extreme would be that all working phases are performed for the first group of questionnaires before starting with the second group. The concentration on a single well-defined job increases the quality and quantity of work performed. At the same time an excessive concentration could lead to boredom and a decline in efficiency. Therefore, a certain specialization could be established which would combine the two extreme methods mentioned before, in such a way that the first working phase for the second group of questionnaires coincides with the second working phase for the first group of questionnaires a.s.o. This specialization implies a more complicated organization in the office with due supervision of each working phase and well organized handling of all census material.

8. Traditionally, the order of operations in preparation for data processing is the following: checking-editing-coding. However, because of the fact that nowadays the quality control of entry omissions is generally done during the enumeration stage rather than during the post enumeration (i.e. data-processing) stage and also because many countries are shifting their integral consistency checks to the mechanical processing stage, the separate editing operations have generally lost their importance.

With modern methods of data processing, after the field checks the central office receives the material and starts the coding of census questionnaires perhaps after a sample revision of the material. If mechanical editing is not possible, it is also preferable to establish a combined editing-coding operation, when the processing worker edits and codes the census questionnaires at the same time.

Because of this, the order of treatment in this document is checking-coding-editing.

II. CHECKING

9. The checking of population and housing census questionnaires is generally a multi-stage operation, both with respect to location and content.

10. It is highly desirable that the first stage checking should be done as close to the enumeration date and to the respondent, i.e. to the field, as possible. Once the questionnaires have been forwarded to regional or central processing offices, the money and time needed for checking individual information in the field is much greater than if the checking is carried out at an earlier stage.

In a well organized census operation, however, after this first checking as close as possible to the informant it is desirable to establish a second checking stage, which may control in different ways and in a more specialized form the quantity of census questionnaires. Sometimes this part of the operation is decentralized to regional processing centres, at other times it is done after receiving the census materials in the central census office.

11. In some cases the first checking even forms part of the enumeration stage itself, because the enumerator himself has to check the census questionnaires. This is especially the case with the householder method, where the enumerator always examines each questionnaire for errors or omissions. But generally also with the enumerator method, the enumerator has to make a "self-control" at the end of his enumeration, using his control lists or enumerator's sheet. The field supervisors' work is especially aimed at control during the enumeration stage. Even with the best organized and performed enumeration and enumeration field control, it is necessary to establish a very complete second stage control, either at regional or national levels or even at both. This second stage control can also be based on sample methods.

12. According to its content the checking phase can be divided into two main types:

- (a) checking the coverage;
- (b) checking on entry omissions.

These two types of checking are generally organized separately, but in some cases, e.g. post-enumeration field checks, the two tasks can be combined.

A. Checking the coverage

13. The coverage checking has to be started during the enumeration stage. For this purpose the best method is to prepare, at a preliminary stage the so-called census control lists. They are used to help the enumerator to cover the population house by house and living quarter by living quarter. They are also helpful as a check after the enumeration, to aid the enumerator to verify whether every given address was enumerated or not.

The same kind of local records, prepared by enumeration areas, can be used by field supervisors and/or local census officers for detecting coverage errors during the enumeration or immediately thereafter, by comparing the completed questionnaires with the available records.

When differences occur between the questionnaires and the control lists these differences may - in both cases - indicate errors in the enumeration as well as errors in the preliminary records. Therefore the differences found should be examined on the basis of new visits, before corrections, in one or both can be introduced. If checking is performed in this way, coverage errors are reduced to negligible proportions.^{2/}

A great help can also be obtained from the census and sketch maps, where the completeness of the enumeration can be examined and also possible double-enumeration by two or more enumerators. But such a control can only be done by the local census office, immediately after all enumerators finish their work, or by field supervisors, who control the work of several neighbouring enumerators. Sometimes the supervisors are obliged to make a certain number of revisits (perhaps on a sample basis), to detect the coverage errors as well as entry omissions.

^{2/} The preparation and use of census control lists is discussed in document ST/ECLA/Conf.32/L.13 of this Seminar.

14. It is part of the coverage check to ensure that census schedules of all enumeration areas are received from enumerators. For this purpose control sheets should be prepared, in which a check should be made to see whether all questionnaires are sent and also whether all enumerators' material are received.^{3/} This control has to be done at all levels (locally, regionally and nationally in the central office), and each level should ensure that each unit of enumeration in a given administrative territory is included in the material when it is sent to the higher level. The material should be remitted with a forwarding statement showing the complete list of those units and the particulars of filled enumeration schedule for each bundle. This list serves then for control at the following administrative level and at last at the central office.

For a careful check, adequate staff should be available in each territorial level, especially in the regional processing centres. The first task is to see that all the schedules listed in the forwarding statement have been received. A double check should be made to see that every unit of territory is fully accounted for. Some checking at random should also be made with reference to the location code numbers marked on each questionnaire to see if the questionnaires are placed in their proper places and not mixed up with those of a different territorial unit.

15. If no current population records are available, early coverage checking may be performed on the basis of probability, by comparing the numerical results of the enumeration with estimates on the basis, for example, of current statistics. This checking should not be postponed till after the questionnaires are received and handled by the processing office, but be executed on the basis of control forms sent to the processing office well in advance of the expected receipt of the questionnaires. These control forms could be made by the enumerators immediately after they have finished their enumeration job. The forms contain a numerical summary of the enumerators' accounts and serve principally for the publication of preliminary census results. At the

^{3/} The control forms indicated, for example, the number and type of documents being sent and received and the dates the documents were expected and actually received.

processing office the data are immediately aggregated for those areas for which estimates have been prepared in advance. If the comparison of the data thus obtained with the estimates available fall outside certain limits, the forwarding of the questionnaires of the whole area is held up till a decision has been reached about which of the two figures is better, the error is then traced and corrected. Such coverage checking should be organized in so that it can be performed quickly and efficiently. Nevertheless it can be established only in those countries or in such part of the countries, where an acceptable civil registration or other administrative recording system exist, and where on the basis of these registrations, the census office could establish local population estimations.

16. When no early coverage checking is envisaged, e.g. because no local records exist, or because estimates of expected local returns cannot be made on a reliable basis, or because the time and money needed for a reliable early check is deemed to be too high, coverage checks should be made afterwards on a sample basis. These so-called post-enumeration checks may also be effective combined with the early coverage method described in paragraph 13.^{4/} However, post-enumeration checks have the disadvantage that they only lead to the detection of the factual error in coverage, but provide no means for correcting this error, which the previous procedure described does.

B. Checking on entry omissions

17. This part of checking is really a preparation for the next phase of data processing, i.e. for the coding. In some cases it is difficult to divide this part of the operation from the editing operations, where also a pre-coding work is done by eliminating omissions, biases and errors which might be present in the questionnaires and prevent errors from entering any subsequent processing stage.

^{4/} For details of the use of sampling related to post-enumeration field checks see document ST/ECLA/Conf.32/L.12 of this Seminar.

These two stages have really more or less the same task; however, it is possible to divide them, not so much by content but by the place and time of implementation. Generally speaking, the first phase, which is called checking on entry errors, should be carried out as far as possible during or immediately after the enumeration, and its main purpose is to produce fully completed census questionnaires. The editing phase is essentially centralized work, and its main purpose is not to complete the questionnaires but to detect and correct all errors, especially with regard to the internal consistency of the entries.

18. For an easy and effective check on entry omissions the design of the census questionnaire is obviously of decisive importance. If the questions to be answered by certain of the respondents only are scattered throughout the questionnaire, the enumerator who is responsible for filling in the questionnaire may easily overlook some of them. It also becomes more difficult to check the entries to see whether they are complete. Questionnaires should therefore be designed in such a way that the person responsible for filling them in has no problems as to whether or not certain entries should be made. The simpler and clearer the design in this respect, the less likely is it that there will be entry omissions.

As regards checking, the guiding principle in preparing questionnaires is to separate the questions that relate to different population groups. It will therefore be necessary to establish a group of questions to be answered by the total population - the so-called general topics - and then to include other special groups of questions referring only to special sectors of the population, e.g. educational characteristics, fertility characteristics, and type of economic activity, which are applicable only to persons over a given age, or detailed economic characteristics, which are to be answered by the economically active only. Some part of the questionnaire could be set aside for indicating clearly which population is concerned through the checking, it can easily be seen from the other control questions, whether a particular person should or should not answer a specific part of the questionnaire.^{5/}

^{5/} For details of questionnaire preparation see document ST/ECLA/Conf.32/L.6 of this Seminar.

19. Checking of entry omission can only be done locally, since only a part of the omissions can be completed through other questions, and there are numerous entry omissions which can only be corrected by revisiting the household.

Therefore the best checking method is to establish a field supervisor organization, whose main obligation is to check the questionnaires completed by the enumerators. If the errors could be corrected through other questions, that method could be used, but if the omission is of such a nature that the gap cannot be filled with the logical aid of other answers, the supervisor himself or the enumerator must revisit the household and obtain the missing information.

Another good method is for the supervisors to make all the enumerators in their charge meet at a common place and exchange questionnaires with one another for a complete check. Of course, this can only be done after the enumeration is completed. They must be clearly instructed about the kind of internal consistency there should be between answers to different questions, and how to set right any omissions or discrepancies. Several omissions might be the result of mere oversight and can be set right then and there by the enumerator who fills in the questionnaire. Any corrections to be made should be effected with the approval of the supervisor and by the original enumerator who filled in the questionnaire. This checking and editing at very early stages close to the source of data will be a great help in ensuring high quality census results, and obviating considerable difficulty at later stages of processing.

20. For this kind of checking it is necessary to establish general standards for instructing all supervisors, as to the corrections or completions that are permissible in different cases. It has to be stressed clearly that all procedures must be very detailed, because over-checking and over-assignment for missing characteristics can also lead to significant errors and disadvantages.

The entry omissions must therefore be divided into two groups:

(a) When there is a possibility of completing the entries with the help of other information. In these cases it is necessary to give fairly strict instructions. A scheme for typical omissions and discrepancies and the method of correcting them is given in Annex I of this document.

/(b) When

(b) When it is impossible to correct the questionnaire or fill in the missing data (e.g., on the age of a married person, or the occupation of an economically active person, etc.). In this case, it must be specified when it is necessary to revisit a household, or when the questionnaire can be accepted as it is, and a coding of "not stated" used later on.

21. Naturally, after the census material has been received in the census processing office, another checking phase can be organized to detect and correct entry omissions. This phase is, however, rather a part of the editing, or, in some cases, of the coding, when these are combined. Nevertheless, if, during the enumeration or later in the field it is not found possible to undertake a proper check, it may be necessary to establish a separate checking phase before the editing and coding. This could be done either with all the questionnaires or on a sampling basis. The difference between this type of checking and the checking which can be done in the local office during or after enumeration is that, in this case, it is generally impossible to complete the questionnaires without the missing information, which cannot be filled in from other data given, because there is no possibility of revisiting households. Therefore this form of checking (or may best be called editing) is limited to the correction of entry omissions which belong to type (a) mentioned in paragraph 19.

III. CODING

A. Types of codes

22. Coding simply involves the replacement of the answers to census questionnaires by a numerical code, which is suitable for transference to a punch card or directly to a magnetic tape. The codes used in census data processing can be divided into two different groups:

(a) Simple codes - which are determined by previously given answers in the census questionnaires, when there are only a limited number of possible answers to the questions,

(b) Complex codes - which are mostly groups with a wide variance of possible answers, as it is useless or impossible to give each answer its own separate code. Therefore in these cases, the possible answers have to be classified beforehand in to pre-determined groups of answers, and the answer to the relevant question can be coded only after the group is found in which it is classified.

23. On the preparation of the census questionnaire, it is necessary to know which codes will be used for the data processing. With this knowledge it is possible to determine which are the simple codes, i.e. which of them have a relatively limited variance. This means that the answers can be given in the census questionnaire. "Pre-coding" can be used for these answers, i.e. in the questionnaire not only the given answers but also the given code numbers are printed. As a result of this procedure, the coding is already completed during the enumeration. It also means that the simple codes can be coded during the enumeration phase, and in the post-enumeration phase need only be checked manually or even automatically, during the data processing phase. This method would save a large number of coding operations, and would not be very complicated to use during the processing phase either, when because of its coverage, it would still represent a relatively large part of all coding operations, sometimes half or more of all the coding time or coding costs.

/24. There

24. There are census questions which, in most cases, can be coded easily by the enumerators, but less frequently need complicated coding operations. With good questionnaire preparation it is possible to establish a mixed coding system to take advantage of this. This can be done by sorting the answers, and by putting in a special general question where the answer is "Yes" or "No". The answer to the general question can be easily coded by the enumerator. The next group of more complicated questions are put only to those persons who gave the less common answer to the special general question. Their answers will then be coded at the data processing phase.

This kind of method is suitable mainly for geographical characteristics, e.g., usual place of residence or place of birth or place of previous residence or place of work, when it has first been asked whether these places are the same as the place where the person was found at the time of the census. If so, the coding is simple, but, if not, the different geographical place must be codified afterwards during the data processing stage.

25. As a rule, it is only when the enumerator method is used that pre-coding can be done during the enumeration stage. But some countries - especially those with a relatively high cultural standard - have also successfully used pre-coding in conjunction with the householder method. In this case, however, the number of pre-coded questions is more limited, because it is impossible to supply the householder with a full list of simple codes so that he can enter the applicable codes on the questionnaire instead of the answers. Nevertheless, mainly for the purposes of subsequent checking, it is preferable to prepare the questionnaire in such a way that the numerical codes are given together with the written answers (e.g., Sex: Male /1/ - Female /1/), whether the householder or the canvasser method is used.

26. It is necessary to reduce the possibilities of having the simple code applied during the enumeration phase to questions with only a limited number of possible answers, that will not be biased by the introduction of the element of choice. In the case of questions which

/may be

may be answered in so many ways that a list of the possible answers would either be incomplete or become too lengthy, it is better not to use that method of coding, even if the nature of the question is very simple (e.g., if it is used, there would be a very long list of countries of birth for the foreign-born population). This would make it difficult, if not impossible, for the householder or enumerator to look up the code for the right answer.

27. In Annex II a list is given of the codes used for the population and housing censuses, divided into the above-mentioned two main groups. Generally speaking, all codes which refer to the general personal characteristics of the population can be given during the enumeration. The geographic characteristics, as mentioned in Para. 25, are those which can be coded partly during enumeration, and partly during the processing stage. The educational and economic characteristics of the population generally need more complicated codes, and therefore are coded during the processing stage. Naturally some of these characteristics have also a simple character, and can be coded during the enumeration phase, e.g. literacy, and type of activity, but because these two groups of topics consist really of a separate unit, it is preferable for every question in these two parts of the questionnaire to be coded by well-instructed and centrally-directed coders in the processing office. This is also the case for codes of household and family characteristics. The information on relationships to the head of the household and the head of the family could be coded during the enumeration phase, but because these have a close connexion with the codes of household and family composition, which are of a more complicated nature, it is preferable to code them during the data processing phase, especially in the case of multi-family households. Housing characteristics are generally simpler; therefore they are usually all pre-coded and do not need any further coding during processing - except perhaps general control, especially on the data for type of building and type of living quarters.

28. The coding of complicated topics is generally done manually by officials of the census processing office. There is, however, a possibility of automating coding operations if reading devices can be used that are designed to read machine or handwriting. Although such devices are being developed, it is doubtful whether they will be generally available for use during the 1970 censuses.

29. The coding of more complicated criteria must also be made as simple as possible. A full systematic classification therefore has to be prepared for all topics. It is preferable to put a separate question on each criterion and to give the answer to each of these questions its own code. In some cases, however, coding is based on more than one question, for purposes of economy. For instance, it may happen that several questions have only two answers. If punch cards are used, it would be a waste to reserve a separate column for each of these questions so long as each column is able to hold ten or even twelve possibilities. In these cases the codes may be combined in such a way that one column holds the answers to two or three questions. This method should be used if really necessary when no further columns are available on a punch card.

When full systematic classifications cannot be established, the coding has to be done on the basis of semi-systematic classification. Hand coding is preferable for these topics based on single questions rather than more than one.

30. With manual coding it is possible for the coders to specialize so that one group of coders works on a given set of topics only. In this respect, it is customary to establish a group of coders for general topics who control the general codes made during the enumeration phase and perhaps code household family characteristics; another group of coders for geographical characteristics; a third for coding educational characteristics; and a fourth for economic characteristics. The work can also be divided in another way, so that established, one for general, geographic and educational characteristics,

/and another

and another for economic characteristics. The coding or control of housing characteristics is generally undertaken during another phase of the work, sometimes together with the coding of household topics.

The division of work is also sometimes based on whether sampling techniques are used during the data processing operations. Two coding phases could also be organized by this method: one for the control of the 100 per cent general topics and one for the more complicated sample topics.

31. For each stage of complex coding it is necessary to prepare special instructions containing, besides the theoretical part on the coding technique of the given criteria, a systematic and alphabetical part dealing with the possible answers and the codes to be used. In general, the following instructions have to be prepared for census purposes:

(a) List of administrative and geographical units of the country, marking the minor and major civil divisions to which they refer. This list should contain, not only the names of the administrative units of the country, but those of all geographical places of localities, because sometimes those are the only ones given in the questionnaire in reply to a geographical question.

(b) List of the main types of educational institutions in the country. The codes are based on the level of education. In some cases individual establishments have to be listed with their respective code, as the name alone does not indicate the level of education given. If educational qualifications are also investigated, they will have to be listed, possibly together with the list of establishments providing the given qualifications. If the educational system is undergoing important structural changes, the list must contain the names and codes of both old and new educational establishments.^{6/}

^{6/} For the preparation of those lists the recommendations of the United Nations Educational, Scientific and Cultural Organization (UNESCO) are very helpful, e.g., Recommendation Concerning the International Standardization of Educational Statistics (Paris, 3 Dec. 1958) and International Standard Classification of Education (ISCED) - Fourth Draft (Paris, 21 Feb. 1968).

(c) Classification of occupations - listing all occupations existing in the country in alphabetical order, with the codes of given occupation. This classification can be combined with that of the possible socio-economic status of the given occupation. The latest edition of the International Standard Classification of Occupation (ISCO), issued by the International Labour Office,^{7/} recommended as a basis for the preparation of the national classification.

(d) Classification of economic activity - listing all industries existing in the country also in alphabetical order. This should be based on the International Standard Industrial Classification of all Economic Activities (ISIC) most recently approved by the United Nations.^{8/}

In some countries not only the industry is asked for, i.e., the economic activity of the establishment in which the person is working, but also the name of the establishment. In that case another list, the so-called Company Name List, must also be prepared. This list should contain the names of the larger establishments with their main economic activity.

B. Form of coding

32. The format of the questionnaire will have been developed in each country to suit the system of data processing to be used. The place and type of coding operation vary considerably, but the methods used can be divided into the following groups:

(a) When simple coding is used the answers are pre-coded, and all countries are coded at the time the answers are given. The format can be varied, depending on the following working phases - punching or

^{7/} The latest revised version of ISCO was adopted in 1966 by the Eleventh International Conference of Labour Statisticians. The publication of the English and Spanish versions is scheduled for 1968.

^{8/} The latest version of ISIC was adopted by the United Nations Statistical Commission at its fifteenth session and is being prepared for the press.

the use of an optical sensing device. For example, the format used might be one of the following for the question of marital status, if the given person is married:

Example 1.

Marital status 1. single - 2. married - 3. widowed - 4. divorced
 5. separated

Example 2.

Marital status single (1) Leave blank
 married (2) **2**
 widowed (3)
 divorced (4)
 separated (5)

Example 3.

Marital status single - married - widowed - divorced - separated



Example 4.

single - married - widowed - divorced
 separated

Examples 1-2 are used for punching or manual sorting and examples 3-4 for optical sensing devices.

(b) The answers, when manual coding is used during the processing phase, can be coded in a separate space reserved after each question for the use of coding, or even the given code number can be written over the answer, e.g., the form of coding the occupation of the respondent could be one of the following:

Example 5.

Occupation tailor Leave blank
 7 **9** **2**

Example 6.

Occupation tailor 792

Example 7.

Example 7.

Occupation

tailor

For office code

0 o o o
1 o o o
2 o o ●
3 o o o
4 o o o
5 o o o
6 o o o
7 o o o
8 o o o
9 o o o
A o o o

Example 7 is used when optical sensing devices are applied.

33. The general form of coding is in the census questionnaire itself, using one of the methods described in paragraph 33. This is used not only for the basis of card-punching, but also when the questionnaires can be put directly through an optical sensing device. In the latter case either the questionnaire itself can be used, or the forms can be microfilmed and fed into an electric film optical sensing device for extremely fast transfer of data on to a magnetic tape which is then processed through electronic computers for detailed tabulation. The format of the questionnaires has to be framed in such a way that they can be used, after coding, for direct microfilming and reading by a film optical sensing device. The development of various electronic mark sensing devices or film optical devices makes it possible to have the census questionnaires processed very quickly if the answers are given in a suitably coded form, thus saving time and manpower.

Sometimes the questionnaire itself, because of its format or very complicated structure or even because of the decentralization of the operation, cannot be used directly in the later phases of data processing. In this case, the answers or even codes furnished by the householder or enumerator are transcoded on to another suitable form or card in the processing office, which then become the basis of the later processing. This is done by electronic film optical sensing methods, or even by one method when manual sorting is used. These so-called "code-cards" can be prepared together for a larger group of

/persons - especially

persons - especially for further filming - or individually if manual sorting is to be used. The personal cards are suitable only for manual sorting. This method can be used when mechanical tabulation equipment is not available, or as a preliminary to mechanical tabulation, when pre-sorting makes the later phases easier.

C. Quality control

34. Whatever method of data processing is used, the selection of the method to be used for coding is one of the most important parts of census preparation. Every effort must be made to reduce coding time and manpower requirements, but even so, this will be one of the most expensive stages of the census. Some of the enumeration errors can be eliminated to some extent during the coding, but it is also possible to make more errors if coding is badly prepared and carried out. Hence it is necessary to prepare all coding instructions in very detailed form and to pick the best possible staff to do the work. They have to be carefully instructed and supervised during the whole coding phase. If possible the system of payment adopted should take into account not only the volume but also the quality of the coding work done.

35. The coding must be checked by all possible methods of quality control. At the beginning of coding a 100 per cent check generally has to be used to ensure that the coders are familiar with the instructions. From then on during the whole coding phase, quality control can be organized on a sample basis. The best method is not direct checking, i.e. when the supervisor sees only the codes which the coder was given, and checks to see whether they are well-applied or not. With this method there is very often only formal supervision and mistakes may be overlooked. This method is called dependant verification. For example, an evaluation of the 1950 census data in the United States, when this method was used, indicated that a verifier might fail to find as many as half the errors in the work he was checking.

A more developed and more effective method needs separate coding. In a sample of census questionnaires the coding is done by the general coders not on the questionnaire - as in the custom - but on a separate coding card. The supervisor gets the same sample questionnaires but without the code cards, and he also does the coding on a separate card series. So the independently coded cards have to be compared with each other and differences determined between the codes entered by the two coders for the same item. The differences have to be analysed, and on that basis, it is possible to assess the quality of the coders and of the census enumeration itself.

In addition to this code control, the questionnaires themselves naturally have to be coded: this coding could also form part of the quality control. In this case the matching is based on three different coders' entries. Through the determination of these three codes, the following groups can be established with the errors defined in terms of the "decision of the majority":

- (a) where all three of the coders agree on a code, all three are considered to be correct;
- (b) if two of the coders agree and the third differs, the two in agreement are considered to be correct;
- (c) cases in which all three coders disagree are excluded from the computation of errors, but the proportion of such cases is very small.

IV. EDITING

36. As was mentioned above, most of the traditional editing could be done during the checking in the enumeration phase and the remaining "pre-editing" work could also be carried out in conjunction with the coding. However, in its modern form, the editing is done after the coding, and not manually, but mechanically, either by unit record or electronic computer methods.

37. With such methods, the editing has two main tasks, as was mentioned in relation to the checking of census questionnaires:

- (a) Editing of entry omissions;
- (b) Control of inconsistencies in census data,

Both tasks are based on the informations already coded and punched or entered on magnetic tape by the computer techniques.

A. Editing of entry omissions

38. Theoretically the editing of entry omissions is based on the same principles as those discussed before and presented in Annex 1. The difference is that in this case, the basis consists of all the code-numbers which refer to an "unknown" or "not stated" answer. In some countries such codes and all tabulations contain a greater or smaller number of persons of "unknown" or "not stated" age, occupation, industry, etc. In other countries efforts are made to eliminate this flaw in the census tabulations, and in place of the missing data "invented" information is assigned. Naturally this assignment could be done manually during the revision phase, with the help of some assignment tables or diagrams, as, for instance, when the missing age data are assigned by this method. But it is much easier and safer to do the assignment after the coding by electronic data-processing machines, which can be given the necessary programming regarding the quantity of missing informations, as this method can be used only if the proportion of "unknown" codes is kept within acceptable limits.

39. The assignment of missing characteristics should be based on the distribution of characteristics for the appropriate population subgroup. This method could be illustrated from the 1960 census in the United States as regards the procedure used in the assignment of unknown ages, by computer. This was carried out in the following steps:

(a) The computer stored the reported age of a person in a subgroup of the population, classified by sex, colour and race, household relationship, and marital status;

(b) This stored age was retained in the computer only until data for another person in the same group - i.e. another person of the same sex, colour or race, household relationship, and marital status - were processed through the computer. The new reported age then replaced the preceding one;

(c) Whenever there was no entry for the age of a person belonging to this subgroup, the current stored age was assigned.

This procedure ensured that the distribution of ages assigned by the computer for persons of a given set of characteristics would correspond closely to the reported age distribution of such persons in the current census.^{9/}

Naturally this procedure can be used not only for allocating missing data on age, but any other missing information. Nevertheless there are some topics for which it is preferable not to use this method, and instead to leave the data as "not stated". So for missing entries on geographic characteristics, it is best not to make an allocation. It is also preferable that only one unknown characteristic should be assigned to each person, but in some cases it is necessary to allocate all the characteristics.

^{9/} Procedural Report on the 1960 Censuses of Population and Housing, United States, Department of Commerce, Bureau of Censuses, Working Paper N° 16, Washington, D.C.

B. Editing of inconsistencies

40. Besides allocating of data, automatic editing also serves to eliminate internal inconsistencies from the individual data series. Ideally, these internal consistency checks should be done when the documents are being "read" by automatic devices. This, however, is only possible if the programme for checking is short enough to be worked through before the "reading" is finished. This would make it possible to punch only those documents that are internally consistent and to reject the inconsistent ones before they are punched. The rejected documents would be re-offered to the automatic punching device after being corrected. They could be corrected either during the "reading" phase, or later, when the information is on the punch-card or the magnetic tape, by two methods:

(a) Automatically, when the computer itself corrects the wrong information on the basis of probability;

(b) The rejected questionnaire would be re-examined and coded again by specialized coders. This could also be done on the basis of probability, or on the basis of the original document, which may take longer, because the questionnaire has to be found.

41. Any method of checking can be used for consistency checks; a well-established programme has to be worked out for all inconsistencies which may appear in the census code series. This programme can divide the combinations of codes into three groups; naturally the combinations can be based not only on two but on three or more logically connected topics:

(a) A possible combination, in which case the codes are acceptable, i.e., no replacement has to be made;

(b) An impossible combination, in which case one or more of the original codes have to be replaced; e.g., if the combination is age and marital status and the original codes show a person of 10 years old who is widowed;

(c) An uncertain combination, which probably contains an error, but is theoretically possible, as when a person is 13 years old and married.

In the case of impossible combinations it is necessary to replace the data. In the case of uncertain combinations the data can be left; sometimes - especially if such cases are relatively frequent - they must be checked against the original material, but must never be mechanically replaced, 42. For the replacement of the impossible combinations, a very detailed programme has to be worked out, also based on probability. The data used for this purpose could be based on the technique described in para. 40 for the allocation of omitted "unknown" data. But in this case a priority order has to be established generally, i.e. in each combination it has to be decided which data are considered acceptable and which have to be changed; e.g., in the case of combinations of age and marital status the age can be always accepted as fixed while the marital status is replaced on the basis of probability tables or classified tables of the given subgroup of population. Naturally the result will be much more exact if both elements are replaced in turn, even if, in this case, these are not the only elements forming the basis of consideration for selection. The main principle is to eliminate the possibilities of over-editing, which is apt to cause more errors than are contained in the original material.

It has to be stressed that neither the correction of apparent inconsistencies nor the assignment of missing characteristics should be carried to such a point that failures of enumeration are concealed by over-edited results. A record of editing changes is extremely desirable for this purpose as an aid to the proper interpretation of tabulated census results.

43. As accurate checking and editing and also codification - particularly when coding is combined with automatic processing procedures - need an enormous amount of complicated instructions and accurate detailed planning, the 1970 population and housing census needs a much longer period of preparation than the previous census. In order to make sure that the procedures followed will be as successful as was expected, it will be necessary to test them on the basis of an experimental census. Without carefully planned data-processing operations, the automation of post-censal work could increase the cost of the operations and the time needed for them.

DRAFT RULES FOR REPLACE ENTRY OMISSIONS FROM THE POPULATION CENSUS QUESTIONNAIRE

Missing information from the census questionnaire	Information (s) which help to replace the missing data			If there is no other indication: the missing data is considered to be decided by
	Primary data	Secondary data	Other data in the questionnaire	
	Which are given in the column of the given person			
(1)	(2)	(3)	(4)	(5)
1. Place where found at the time of the census	-	-	Census control list Neighbouring questionnaires	-
2. Place of usual residence	-	-	Data of other household member	Place of enumeration
3. Place of previous residence	-	-	" " "	Place of birth or of enumeration
4. Duration of residence	-	-	" " "	No migration
5. Place of birth	-	-	" " "	Place of previous residence or of enumeration
6. Relationship to the head of household	-	Name Sex Age Marital status	Data of other household members (relationship)	Same name: other related Different name: not related
7. Sex	First name	Relationship to head of household Occupation	-	-
8. Age	-	-	-	Not stated
9. Marital status	-	Sex Age Relationship to head of household	Data of other members	Below given age: single Over given age: Not stated
10. Literacy	-	Age Educational attainment School attendance Occupation	-	Below given age: illiterate Over given age: Not stated
11. Educational attainment	Age School attendance	Literacy	-	Below given age: no education Over given age: Not stated
12. School attendance	Age Educational attainment Type of activity (functional category: whether student)	Literacy	-	Below given age: not attending In a given age (e.g. 5-24): not stated Over given age: not attending
13. Type of activity a) general (active-inactive) b) employed-unemployed (actives) c) functional categories (inactives)	Sex Age Direct question on seeking work Sex Age School attendance	School attendance Occupation Functional categories Relationship to head	Data of other members	Below given age: not active Active age: Not stated Over active age: not active Not stated
14) Occupation	-	Industry status	-	Not classifiable
15) Industry	-	Occupation status	-	Not classifiable
16) Occupational status	Occupation	Industry	-	Not classifiable
17) Children born alive	Age Children living	Marital status Relationship to head	Data of other members (number of "children")	Below given age: 0 Over given age: Not stated
18. Children living	Age Children born	Relationship to head	"	Below given age: 0 Over given age: Not stated

General notes

1. In order to check and replace entries (when missing) it is necessary to look for other information on the same person, i.e. for answers given to other questions which can help to provide data to fill in the omitted or unacceptable entries. Such data can directly provide the information required ("primary data") or can be of help only in certain circumstances ("secondary data"). In other cases, the replacement can be done with the help of information given on another person in the same household (especially if household questionnaires are used).
2. In some other cases, information from other questionnaires could also be used to correct the questionnaire and also to supply the missing information. This is generally the case with questions on living quarters, when missing information can be easily obtained from the questionnaires on neighbouring living quarters, especially if they are in the same building, and even if they are located in a separate (one-dwelling) building.
3. If no other indications can be found, there are possibilities of replacing the missing data mechanically (as mentioned in Column 5). But in other cases mechanical replacement is not allowed (or only by the method of assignment, as explained in this document).

Annex II

POSSIBILITIES OF CODING OF POPULATION CENSUS DATA

Characteristics	Coded during		Number of digits used.	Special instruction ^{a/}
	Enumeration	Processing		
<u>Geographic characteristics</u>				
1. Place where found at the time of the census	x	-	6	(A)
2. Place of usual residence	If same as 1	If different than 1	6	A
3. Place of previous residence	"	"	2	A
4. Duration of residence	"	"	1	-
5. Place of birth	"	"	4	A
6. Urban-rural	x	-	1	(A)
7. Locality group	x	-	1	(A)
<u>Household characteristics</u>				
8. Relationship to head of household	x	-	1	-
9. Household composition	-	x	2	-
<u>Personal characteristics</u>				
10. Sex	x	-	1	-
11. Age (or year of birth)	x	-	2	-
12. Marital status	x	-	1	-
<u>Educational characteristics</u>				
13. Literacy	x	-	1	-
14. Educational attainment	-	x	2	B
15. School attendance	-	x	2	B
<u>Fertility</u>				
16. Children born alive	x	-	1	-
17. Children living	x	-	1	-
<u>Economic characteristics</u>				
18. Type of activity	General	x	1	-
	Employment	x	1	-
	Functional categ.	x	1	-
19. Occupation	-	x	3	C
20. Industry	-	x	3	D
21. Occupational status	-	x	1	C

^{a/} Special instructions: A: List of territorial units - B: List of educational institutions - C: Occupational classification - D: Classification of economic activity (and/or Company Name List).

