CODING OCCUPATION AND INDUSTRY

PREFACE

This working paper covers issues which must be addressed when questions related to ‘occupation’ and ‘industry’ are to be included in a population census. Preparations for coding of ‘industry’ and ‘occupation’ responses are discussed after outlining the objectives and main strategic choices to be made. Specific operational and organizational factors determining the effectiveness and success of the coding operation are discussed. The development of the major tools, the coding indexes, and how to use them effectively, is described. Following a discussion of a first draft by a meeting of experts in May 1998 this text was prepared to serve as part 3 of an UN/ILO publication: *Guide for the collection of economic characteristics*. Part III of *Handbook of population and housing censuses*. Studies in Methods. Series F, No. 54. It is being issued in this form because the publication of that document has been delayed. The text of part 2 has also been issued as *STAT Working Paper 2001-1*, see *Gilbert, 2001* . Comments, questions and suggestions concerning the contents of this Working Paper may be addressed to: Bureau of Statistics; International Labour Office, CH-1211 GENEVE 22. Fax: + 41 22 799 6957. E-mail: stat@ilo.org.

Table of contents

1 Objectives

1. When coding the information given by a respondent about the place of work (i.e. 'industry') and type of work (i.e. 'occupation'), the main aim is to determine and record correctly to which group in the respective classifications the job belongs. It is important to retain as much as possible of the information contained in the responses given. This task has to be completed within an overall processing plan for the census. There are normally a pre-specified timetable and cost limit, to minimize costs given the specified data requirements. Considerations of expense and quality of the resulting statistics have to be balanced when making the preparations for this part of the census operations.

2 Selecting the classifications to use and the necessary tools for their use

2. Most of the discussion in this paper assumes that there exists one national standard classification for 'industry' and one standard classification for 'occupation'. However, this cannot be taken for granted: In some countries there may exist more than one "national" classification for either 'industry' or 'occupation', and in other countries there may not have been the capacity to develop one or both classifications. In either situation a dialogue with possible users of the census results should include a discussion of their needs for statistics with 'industry' and 'occupation' specifications, and to what extent the possible options can satisfy them:

(a) In a country where there are several well-developed classifications, each preferred by different set of users, the census planners will have to investigate with the users and custodians of the different classifications the following issues: (i) whether one of the classifications can serve as a "reasonable approximation" for the others; and (ii) whether it may be possible to design the coding process to serve all relevant alternative classifications reasonably well. (One way to implement (ii) is described below in section *5.9 Coding to more than one classification*.)

(b) In countries where there are no satisfactory classifications, census planners must choose between developing classifications for the census, or using, with only minor modifications, the models which would have been chosen if one was developing national standard classifications. The respective international classifications are often the preferred models.

In either situation (a) or (b), the census planners have to ensure that (updated versions of) the classification to be used will be ready in time for the census. Each classification should cover all types of activities and work situations of the population, and the needs of users of census results should be adequately catered for by the classification structures. Adequate data collection and coding tools must be developed in time for the census. Priority should be given to the development of coding tools which can be used with whatever ‘industry’ and ‘occupation’ classification chosen. The development of national standard classifications are major undertakings which should not be done as part of the census preparations if it means that the development of good data collection and coding tools will be adversely affected, e.g. because of insufficient resources to carry out both tasks well.

3 Strategic coding and processing options

3. Given the objectives and limitations mentioned above, to develop a coding and processing strategy census planners need to consider:

(a) the existing data processing infrastructure and capacity of the organization responsible for the census;
(b) the type and format of the information to be processed;
(c) the number of census questionnaires to be processed and the speed with which the processing needs to be done;
(d) that processing the ‘industry’ and ‘occupation’ responses is part of the total data processing task for the census.

The precise impact of these factors on census operations and the quality of the resulting statistics will depend on the strategic choices made for coding and processing of the census forms. These issues are discussed in the following sections.

3.1 Processing all cases or a sample only

4. The coding of ‘industry’ and ’occupation’ is typically among the most expensive and time-consuming operations in the processing of a census. To reduce costs, to ease the management and quality control of coding and to enable results to be produced early, census planners have to consider whether to code this information only for a sample.

5. To code ‘occupation’ and ‘industry’ for a sample of census respondents can be implemented by asking the questions on ‘industry’ and ‘occupation’ only of those included in the sample. However, this makes it necessary to use two different questionnaires, one ‘long’ and one ‘short’. It may be difficult to organize this correctly during the field work. An alternative is to collect the information from everyone, but do the coding only for a sample of respondents. Either solution will reduce the cost of coding almost proportionately. However, the cost of preparing for the coding operation will not be reduced. In addition, coding only a sample means the

introduction of sampling imprecisions (sometimes incorrectly and confusingly referred to as "sampling errors"). This will be an important concern when estimates are to be produced for small population groups or small geographic areas. As the provision of small area and small group data on a nationally consistent basis is seen as one of the major functions of the census in most countries, many users may see the availability of industry and occupational information for a sample only as completely against the objective of the census. This viewpoint is especially pertinent if there exists a regularly conducted, e.g. annual, labour force survey which already provides national statistics by these variables, on a sample basis. Coding 'industry' and 'occupation' for a sample only in the census will also mean that sampling imprecision will make comparisons between groups and over time more difficult or uncertain.

6. If the decision is made to collect and/or code 'industry' and 'occupation' information only for a sample of the population, then attention must be given to the design of the sample. The sampling fraction should in principle be determined by balancing the precision required for the smallest aggregates of the population for which separate estimates are to be produced, against the saving of cost and time resulting from reducing the amount of coding. The sampling has often been done by selecting all households in a whole enumeration area to be in the sample, as this makes it possible to send whole bundles of questionnaire into the same processing stream. However, this may result in very high clustering of respondents to particular occupations and industries, and contribute to quite significant loss of precision for many regional and national estimates. Using total counts for all localities smaller than a threshold and different sampling fractions for larger areas may provide adequate data quality for those needing such statistics for localities. However, this procedure will increase the complexity of tabulations and the imprecision at the regional and national level, compared to the use of a uniform sampling fraction for the whole country.

3.2 Field or office coding?

7. The choice between the following strategies for coding has important consequences for the costs, quality and control of the coding process:

(a) the respondent coding himself/herself to a predefined group;
(b) coding by the interviewer/enumerator during the interview or before the questionnaire is forwarded for further processing;
(c) coding by specially trained coders as part of the data entry and consistency control of the questionnaire.

3.2.1 Coding by the respondent

8. This means that the respondent is requested to place his/her job in one of a set of predefined groups which will be read to him/her, or be presented in written form on the questionnaire or on a card shown by the interviewer/enumerator. The main advantage is that it is the least expensive of the possible coding procedures. The main disadvantage is that the quality of the resulting statistics will be low, for two reasons:

(a) It is difficult to assure that the respondents will understand the pre-defined groups and relate them correctly to their own jobs. The intended content of each pre-defined group has to be described with a limited number of words, normally in the form of a group title which a respondent may not easily associate with his/her job or place of work.

(b) Many users of 'industry' and 'occupation' statistics need to have much more detailed distinctions and to work with much more homogeneous groups than those obtainable when using this approach.

However, the cost advantages of this strategy are such that some national statistical agencies have been using it in their census, thus foregoing the census' main advantage and *raison d'être*: the possibility of providing consistent statistics for (relatively) small groups for the whole country as well as for small geographic areas. Before choosing this strategy the census planners should therefore make sure to consult all relevant users of these statistics.

3.2.2 Coding by the interviewer

9. Coding by the interviewers/enumerators can take two forms: Similar to the above strategy is the possibility that the enumerator assigns the 'occupation' and 'industry' responses to pre-coded alternatives during the interview, based on the information received from the respondent to the standard questions. The cost advantages are almost the same as for respondent coding. The main difference is that the interview is likely to take a little longer, because of the need for the enumerator to understand the information received and then 'translate' it to the appropriate groups in the classifications. This procedure should be preferred over the one discussed above if the respondents themselves are not expected to (be able to) read questions and write answers. The consequence of this procedure for the usefulness of the resulting statistics is as described under (b) above. The reliability of the coding may, however, improve compared to coding done by the respondent, because the enumerator may receive much more detailed instructions and training on what types of jobs the different pre-coded groups are supposed to cover and the type of ambiguities which will require probing. However, there is also the danger that the enumerator may misunderstand what the respondent tells him/her and therefore select incorrect groups.

10. The other possibility for coding by the interviewer is that he/she writes down key words of the respondent's answers and then code the responses after the interview, but before the questionnaire is forwarded to the processing centre. The advantage of this procedure over those mentioned above is the possibility which it gives for much more detailed coding and thus much more useful statistics. An interviewer who is coding outside the immediate interview situation may also be given complete coding indexes to assist the coding process (see section 5 below), as well as other coding tools - including the possibility to forward queries to supervisors. An added advantage is that the interviewers often will retain in their memory more details about the response than they were able to write down on paper. Another advantage is that as the interviewers gain experience with coding they will become more aware of the type of information which is needed to code correctly. It may also be possible to save time, costs and operational complications by using interviewers to code 'industry' and 'occupation' when these variables are the only items which otherwise would require office examination, because then the questionnaires can be passed straight from the field operations to computer data entry.

11. The main disadvantage of interviewer coding is that because the interviewers are numerous and geographically scattered, they cannot be given the same amount of training, supervision and support as specialized coders, with negative impacts for coding consistency and reliability. Field coding also sacrifices to a large extent the important advantages of a controlled and supervised coding environment which can provide a direct feedback on coding quality to those who are coding. Such direct feedback can be provided to office coders, through the quality control procedures, but will be much more difficult to provide to the interviewers: Field operations are likely to have been finished by the time serious problems are discovered, and the field supervisors will normally not be equipped to give quality control of coding high priority, or be trained to serve as first line query resolvers. They should, however, be required to control whether the 'industry' and 'occupation' questions have been adequately answered.

12. The arguments related to the use of field coding close to the respondents, instead of specialized coders close to the processing of the census forms, are closely linked to whether or not it is possible to have a permanent field staff which cost effectively can be trained in coding and which can accumulate experience. This means that field coding may be a realistic option for continuous labour force and similar surveys, as well as for local administrative offices, **but that *ad hoc* and much larger scale operations like population censuses should normally leave the coding of 'industry' and 'occupation' to the central processing operation.**

3.2.3 Centralized coding

13. The most common solution for population censuses, and the one which will be assumed for most of the remaining sections of this chapter, is to incorporate coding with the rest of the processing operations for the census. Depending to a large extent on organizational

considerations census processing and coding will be located in one national center or in a limited number of regional centers. The number of processing centers and their locations will be closely linked to the size of the census operation, i.e. the size of the country. (The processing of a census creates a large, but temporary, demand for suitable staff and premises which in large countries can only be met by having several locations.) The main advantage of centralized coding is that the work can be organized to ensure close and fast communication between the regular coders and their supervisors on the one side and the experts coders and the classification experts on the other. This communication is particularly important early in the coding process, when many unexpected situations and cases will emerge. Then queries need to be resolved and communicated quickly to all coding teams, together with the consequent adjustments to the coding instruments. Thus the use of several processing and coding centers may lead to difficulties in maintaining consistency in coding between coders and coding teams, as well as in the application of consistent production and quality control standards more generally. It will be important to establish possibilities for communication by telephone, fax and/or electronic mail between the location of the classification experts and main query resolution team on the one side and the various coding and processing centers on the other.

14. Coders may be entirely specialized on the coding of one variable, or specialized on coding in general [i.e. on ‘industry’, ‘occupation’ and sometimes ‘education and training received’ (‘qualification’)]. It is, however, generally recommended that they carry out the coding as one part of an integrated data entry, coding and data control operation for the census to reduce some of the risk of fatigue due to very monotonous and repetitive tasks. **The use of specialized coding units should be avoided. Coding should be one element in a larger processing task for each operator**. This solution also means that fewer persons will be handling the census forms, and this simplifies the task of controlling the flow of questionnaires, see section 3.10 below.

4 Planning and organizing coding operations

15. Assembling the right resources for coding ‘industry’ and ‘occupation’ in the right place(s) at the right time, and managing those resources efficiently, are fairly complicated tasks. They require anticipation and co-operation between different parts of the organization and must be well integrated with the rest of the processing. Large volumes of documents and data have to be handled. The interdependence of the different stages of the process means that the penalties for failures of operational or quality control may prove to be heavy: Delays and cost increases as well as reduced quality and credibility of the census results are likely consequences. The requirements for finance, staff, equipment and premises may need to be specified long before the start of the actual processing of census returns, based on inputs from and cooperation between:

(a) Managerial staff involved in the planning and supervision of the coding operation;
(b) Professionals concerned with the design of the classifications and the coding procedures,

the training of coders, the updating of the classifications and the interpretation of results;

(c) Staff responsible for planning and implementation of other aspects of census operations, e.g. questionnaire design, field staff, data entry and tabulations.

4.1 Finance and resources

16. Substantial amounts of money are required to support the processing of a census and these will have to be estimated and provided for under appropriate budgets. Estimates for each part of the processing task often need to be made several years in advance and to be fed into the financial planning and procurement procedures of the responsible agency to ensure that adequate staffing and equipment is available when needed. This demands early decisions about resource requirements and these may in turn precipitate strategic processing decisions which have resource implications (e.g. staff numbers and pay rates, number of processing offices, whether to use computer-assisted techniques). It is important to ensure that financial, resource and operational planning are coordinated, so that the technical assessment of requirements determines bids for resources, rather than *vice versa*. As coding of ‘industry’ and ‘occupation’ will constitute major cost elements for census processing, a clear understanding of the cost of this part of the operation is essential.

4.2 Expertise, experience and rehearsal

17. Processing of each census relies heavily on technical information, expertise and experience gathered at the last similar exercise. Information and experience gained then may be documented in detail, but most of the practical expertise is likely to reside in the heads of a small number of experienced staff. Staffing continuity in key positions is therefore extremely desirable. However, circumstances change and it is not possible to rely entirely on the experience from the last census. The staff working with the *Labour Force Survey* or other surveys where information about ‘industry’ and ‘occupation’ is collected on a regular basis will often have very relevant experience and tools which should be consulted. If and when outside advisers are used it is also extremely important to verify the validity of their experience for the local context and to have the opportunity to modify their estimates of time requirements and costs, based on concrete field experiments.

18. A processing rehearsal is very important as an aid to planning and estimation, and the results of such a rehearsal need to be available at a stage when they can lead to adjustments in the plans for the main operation. Some key planning parameters (e.g. the number of questionnaires to be handled by one coder per day and the number of questionnaires which must be processed per working day in order to finish before the deadline set for the census) can only be reliably estimated from such rehearsals, even with well documented previous experience. However,

certain problems arising from the scale of the full operation - e.g.. its effect on problems of recruiting, maintaining and controlling staff - may be hard to test in advance, as experience shows that performance rates vary significantly over time during the processing period. Coding rates are much lower and query rates are much higher early in the process than later, and the signals from a rehearsal may be more indicative for the early phase of the operation than for the latter. There is also the danger that differences over time can be created by relaxing controls and standards towards the end of the process, beyond what is warranted by the improvements in the coding operation. Such improvements should be expected because of the early improvements in the coding tools used and the on-the-job learning by the coders, supervisors and classification experts. However, after some time few new queries or operational difficulties are seen to arise and managers and data users often take this as evidence that the coding task presents few new problems and is being carried out to high standards of accuracy, reliability and validity. This assumption and situation may, however, be dangerous. It is probable that intractable coding problems may not have been solved, only “dealt with” *ad hoc*. Individual coders also tend to establish 'short-cut' methods which reduce the laboriousness of their task but which may also incorporate errors or unjustified assumptions, if not actual violations of the coding instructions. These unofficial additions to and departures from the specified coding procedures tend to become institutionalized to the point where no distinction is perceived between them and coding rules derived logically from and designed to support the classifications used. Therefore the routine procedures of the coding unit should include specific and properly designed checks to ensure that proper procedures are followed and the required quality of coding achieved.

19. The large coding exercise mounted for national censuses of population has to rely on the recruitment and training of inexperienced coding staff. Using inexperienced staff may delay the acquisition of the bad coding habits described above, but the coding results are likely to be poor unless recruitment, training and coding procedures are well planned, executed and supervised. A particular danger is that, because of resource limitations or practical difficulties at a time of heavy stress in the preparation for the census processing, there may be insufficient time to establish rules and routines and generally 'run in' the coding procedures before production coding has to start. If this happens the organization may initially be temporarily overwhelmed by the sheer volume of documents and data to be handled, with a consequent loss of control and severe fall in standards. Such control may prove difficult to re-capture and reverse later.

4.3 Coding staff

20. In a census operation the large volume of work to be processed within a limited period will require special efforts to recruit and train staff, also for anticipated staff turnover. Thorough enquires and consultation should be undertaken about likely sources of suitable staff recruits, since financial constraints are likely to prevent actual recruitment until the last moment. There may be external pressures to employ particular groups of persons, even when their suitability

cannot be guaranteed. Clear criteria should be defined and applied consistently in the selection of all staff. The terms on which staff are to be employed, including the minimum acceptable level of education, pay rates, grading, disciplinary and hiring/firing rules, need to be carefully defined. It is important to provide adequate time and resources for staff training for both coders and supervisors. It should also be recognized that the specialists, who are to advise the supervisors and to resolve the more difficult queries, normally cannot, and should not, be recruited and trained for the particular census operation. They should be part of the permanent competence of the statistical office.

21. The tasks of the coder are best performed by persons with the following characteristics.

(a) Literate and reasonably intelligent, but not over-eager to display independence of judgement as this may lead them to find the task demeaning or frustrating.
(b) Clerically accurate and careful.
(c) Willing and able to follow detailed instructions conscientiously, without attempting to alter or improve upon them, and prepared to raising queries in cases of genuine doubt.
(d) Honest and trustworthy and thus not likely to falsify or omit procedures in order to reduce the amount of work to be done per case, or for other reasons.
(e) Persistent and willing to work steadily for long periods.
(f) Able to work reasonably rapidly and to maintain a steady level of productivity.

Those responsible for recruiting and selecting coders should have these characteristics in mind. Several of them (e.g. b, c and f) are best assessed through an objective screening test, which may also be applicable to other types of routine clerical tasks. Mistakes at the initial recruiting stage are likely to lead to high staff turnover, and the need to recruit and train more replacements while production coding is in progress.

22. **Proper training of those who carry out the coding is very important**: Coding of ‘industry’ and ‘occupation’ does not require theoretical knowledge. The task is best learned through practical instruction in specific procedures (e.g. document handling routines, use of the coding indexes etc.), interspersed with supervised practice on appropriate, specially designed exercises. Relative slowness in learning need not be a disadvantage if accompanied by good retention of what is learned and the desirable temperamental characteristics. The training period can also be used to identify persons who are unwilling or unable to follow the coding instructions precisely. It is important to remember that training will be needed also for staff recruited to replace those who leave before the end of the operation.

4.4 Coding teams and supervisors

23. It is important to make good estimates of the number of coders required, as well as of the

number of ‘first-line’ supervisors who are needed to control the coding process and of the number of specially trained staff needed to resolve queries.

24. Coding is best organized by allocating coders to teams, each under a “first line” supervisor. The supervisor's role and work tasks need to be carefully specified and are likely to include: controlling work flows; monitoring and maintaining work rates; enforcing work discipline; resolving and recording coding queries: applying quality control procedures; etc. The supervisors need to be trustworthy persons with the necessary intelligence and force of personality to master their duties and control and motivate coders. The need for them to have previous experience in coding operations depends on their role in query resolution: In principle one may organize the coding operation in a way which gives the operational supervisors a very limited role in query resolution. Then it is not essential or necessarily desirable that they should have had prior experience of ‘industry’ and ‘occupation’ coding. However, in most cases it will be preferable to give supervisors the responsibility for first line query resolution because of their close contact with the coders and the longer response time and limited capacity of the classification experts. **Supervisors with responsibility for query resolution should be given a good understanding of and training in the classifications and coding systems.**

25. The number of coders allocated to each supervisor is important. Typical ratios lie between 6 and 12, but the appropriate number needs to be assessed in each case, taking account of the flow of work with which supervisors will be required to cope. Overloading of supervisors will cause bottlenecks and poor staff morale, as well as under-reporting of problems and queries and a reduction in the reliability of coding. Because coding is a fairly monotonous task it is important to ensure that work discipline is maintained and that productivity rates do not fall off. Particular problems may arise when coders expect that it will be difficult to find a new job after the end of the coding operation and who therefore may want to make the work last as long as possible. Special bonuses may be effective in counteracting such tendencies. It may also be possible to retain to the end of the process only those coders who will have a (possibility of) more permanent and long term employment with the statistical service. Their experience will be valuable both for other surveys and for the documentation and explanation of the census procedures to the users of census data as well as to those who in time will prepare for the next census.

4.5 Coding tools

26. It will be necessary to provide appropriate documentation, procedures and training material for both coders and supervisors. The basic tools required will include:

(a) Coding instructions: These should cover all operations which the coder is required to carry out. The procedures and instructions for handling all relevant items and operations should be integrated. The instructions relating to the coding of 'industry' and 'occupation' will need to be particularly clear and specific on: (i) the order in which checking, coding and editing tasks are to be carried out; (ii) the procedure for analyzing the written responses for significant terms; (iii) the use of the coding index; (iv) the circumstances and procedure for using ancillary information; and (v) when to refer a 'difficult to code' reply to a supervisor for query resolution;

(b) Coding index: This is the key coding document through which the written responses are translated into codes. Coders should not be encouraged to interpret the written responses in terms of their own conception of the purpose or criteria of the classifications, but rather to follow in a conscientious way the instructions laid down for consulting the indexes. For these reasons it is essential that the indexes be clearly set out, explicit and easy for coders to use. The instructions and procedures for the use of the coding index need to allow for updating in the light of decisions made in resolving queries and problems which arise and are dealt with in the course of coding. This is discussed in more detail in section 5 below.

(c) Queries: There should be clear instructions on when and how the coders should raise queries, and how to record them and their resolution. Queries are the most useful inputs to both immediate and future work to up-date the coding index and the classifications. It may prove necessary to carry out such updating frequently early in the census processing, if the coding indexes, or even the classifications, prove to be out-of-date or incomplete for other reasons.

(d) Legal and administrative forms: A legally binding undertaking to maintain the confidentiality of census data should be signed by the coders. Other documents used by both coders and supervisors are likely to include forms for recording queries and their resolution; forms for controlling the flow of work and reporting progress; forms for quality monitoring; etc. To ensure that target throughput rates are achieved, the productivity of coders and coding teams needs to be monitored and progress charts maintained. Special measures for motivating coders should be used, e.g. the posting of productivity and error rates for each coding team.

4.6 Coding problems and queries

27. It can be guaranteed that large numbers of detailed queries will be raised in the course of a major coding operation, no matter how carefully coding instructions and the coding indexes have been prepared, This happens mostly because the indexes will prove to be out-of-date or incomplete in some respects. Another reason is that actual responses will be more varied than anticipated by the index constructors, even with the best tested and most carefully designed

questions and instructions to the respondents and enumerators. Any revision of the structures of the classifications since the last census or survey may also lead to new problems, particularly in the treatment of vague and inadequate responses at the borderlines between categories.

28. Evidence of shortcomings of the documentation thrown up in the course of coding, will need to be rapidly and consistently processed and the results should be fed back to the coders and their supervisors in the form of amendments to their tools. Appropriate procedures need to be laid down in advance for reporting and recording queries and the decisions made in resolving them, and for incorporating any consequent amendments into the coding documentation and procedures. The roles of supervisors in processing queries and amendments need to be defined, e.g. how and how frequently they should communicate with the coding experts and how new versions of the tools should be distributed to the coders. Particular care is needed in the coordination of query reporting and documentation of amendments when coding is being carried out in several different locations, e.g. in different provincial or local offices. **All coding sites and teams need to receive information and updated tools as quickly as possible**.

4.7 Quality assessment and quality control

29. Casual observation by supervisors and simple visual checking of coded output do not provide adequate information on quality. Explicit allowances for the resource and time costs of formal quality control need to be built into the processing plan. These costs should cover the establishment and staffing of a quality control unit responsible for acceptance testing of coding during the start up of the coding operation, and for the assessment of the reliability and consistency of the operation as a whole. A procedure for sampling the work of the coders for quality control purposes needs to be defined. The quality control unit needs to be staffed at a level which will enable it to keep pace with the main coding operation. Coding schedules as well as document handling procedures must permit corrective action (e.g. 100 per cent checking) in the case of batches which fail a quality control test.

30. In-built quality control procedures will also be required. It is necessary to design separate procedures to handle: (i) on-line acceptance testing of coders' work; and (ii) overall monitoring and assessment of performance. The aim of acceptance testing is to identify rapidly coders whose performance does not meet the required level of accuracy, so that corrective measures can be taken. The aim of overall monitoring is to estimate average levels of coding accuracy and inter-coder consistency for the entire coding exercise. Note that estimates of coding reliability need to be supplemented by estimates of the validity of coding if a balanced overall assessment of the quality of the statistical output is to be made. Estimates of validity may be obtained from a post enumeration study in which the whole data collection, coding and editing process is repeated for a sample of census returns, see e.g. section II.B of *United Nations (1998)*. On the basis of such quality assessments it may be possible to separate the contributions made to total error and

variance by error/variability in data collection and error/variability in coding.

4.8 Premises, infrastructure and equipment

31. The large volume of census processing requires suitable office space and all the necessary infrastructure for properly supervised clerical operations, as well as for easy movement, storage and retrieval of forms. A special requirement is for security of documents with information about individuals, e.g. the completed census questionnaires. Coding is a monotonous task which makes staff sensitive to the work environment, the functionality and capacity of the desks, chairs, shelves, filing cabinets, as well as the adequacy of lighting, heating and ventilation, and the supply of paper, pencils and other stationery. Neglect of such factors can easily influence the morale of the staff and result in higher than anticipated staff turnover and inadequate attention to work quality and speed. Suitable premises need to be specified, identified, costed, approved and booked well in advance. If coding staff is to operate e.g. computer terminals or optical readers, then special arrangements may be needed to: estimate the requirements; identify suitable and reliable equipment and carry out tests; estimate and provide for capital expenditure and depreciation; provide for replacement in case of breakdowns; etc.

4.9 Handling of documents

32. The coding of ‘occupation’ and ‘industry’ will normally be an integrated part of the total processing of the information on the census questionnaires. Then the main paper handling concerns will be: (i) how to register the reception of the questionnaires; (ii) how to store them while they are being processed; (iii) how to allocate them to staff so that one can control that all questionnaires have been processed once and only once (except for quality control purposes and to correct detected errors); and (iv) when, how and to where to pass on the questionnaires after the coding process has been completed. During this process it should be easy to find individual questionnaires which for some reason need to be rechecked. If each questionnaire has to be handled by more than one person, for example because the coding of different variables are carried out by different persons or because data entry are done by special operators, then the flow of documents must be planned to avoid bottlenecks and the loss of any of the questionnaires. All movements of questionnaires from one location, i.e. workstation or officer, to another and to storage should be carefully recorded.

4.10 Editing of the industry and occupation variables

33. Following the entry of the census returns into machine-readable form the data will normally be tested through a series of computer edit-checks. These are described in more detail

in a separate volume of the UN Census Handbook, see *United Nations (forthcoming)*. Possible test for' industry' and 'occupation' codes are mostly limited to testing for valid codes. There is only limited possibility for testing for logical inconsistencies between 'industry' and 'occupation', or for inconsistencies with other variables. It may be possible to "flag" suspicious combinations of values of certain characteristics, e.g. the "medical doctor" with "primary school" as highest level of educational attainment. It should also be checked that all "employed" persons have been given codes for these variables, or an indication of "missing variable". The same should be done for "unemployed" persons if they are asked about the characteristics of a previous job. (A separate code should be given unemployed persons who did not have a previous job.).

4.11 Use of automatic or computer assisted coding

34. It is only relatively recently that real progress has been made in bringing computers to bear on the tasks carried out, in an inherently slow and laborious way, by census coders. However, the situation is changing rapidly, and systems for automatic (AC) or computer-assisted (CAC) coding of industry and occupations are now used by a number of statistical services. So far the experience has been that the introduction of these methods may have beneficial effects on the consistency of coding, but they do not reduce very dramatically the combined time required for coding and data entry and the elapsed time required to complete the task, unless (i) it is possible to combine the (C)AC systems with automatic (optical) reading of the responses written in free text; or (ii) the processing procedures only require the entry of a few characters from most of the responses. Both features will significantly reduce the task of transcribing the verbatim material into a computer-readable form. Even where operators with appropriate keyboard skills are available such transcription is inevitably slow and error prone, and its role as a limiting factor becomes more obvious as the power, efficiency and speed of the computer hardware and software increase.

35. Some statistical services have reported that the use of optical character reading (OCR) forms, equipment and software contributed significantly to improved efficiency in the processing of their population census in the 1990 round, and in some cases this provided the basis for the introduction of (C)AC systems for coding 'occupation' and/or 'industry'. However, successful OCR operations frequently require the use of special quality paper and ink. Special handling procedures for the questionnaires are needed before, during and after the field work to protect them from humidity, sun and other spoiling influences. This may prove difficult to ensure in many countries. Some interactive CAC systems only require the entry of a limited number of characters from the written response in most cases. This may also lead to a significant reduction in total data entry costs, and may be easier to achieve in many developing countries.

36. So far no workable system has been developed which fully automates the coder's decision-making task. Some AC systems claim to allocate codes automatically to more than 70 per cent of

cases, but the development costs have often been high and there have been problems in making the systems sufficiently 'intelligent' to simulate reliably the performance of trained human coders. The reported error rates for those responses which the systems code automatically are mostly of the same order of magnitude as those of human coders, and these must be considered to be the easy cases. Moreover, the residual need for human intervention in the coding process in a substantial proportion of cases tends to limit the effect of such automation in simplifying, speeding up and reducing the cost of data processing. However, this need does not eliminate the gains to be had by using a (C)AC system, in particular if it is integrated into a data entry and processing system which starts with optical reading of the questionnaires.

37. Although (C)AC systems have the potential to yield significant improvements in the quality of data and to reduce the amount of time which elapses between data collection and data dissemination, due regard must be given to the true costs of implementing the system. Realistic cost estimates include the rate of depreciation of hardware and software, although these may frequently be put to good use after the completion of the census processing, e.g. in regular survey operations. The cost estimates must also include the dependence on specialist programming and systems analysis skills for the implementation of the required software, and the cost of training users to work in the disciplined environment of a computerized processing and coding system. **If it is intended to introduce a (C)AC system, trials of the hardware and software and of the machine/operator interface need to be conducted well in advance. Until the feasibility and operational robustness of the machine-based system have been established, it is prudent to make parallel plans for the use of a manual/clerical system as a fall-back procedure.**

38. To minimize the risks of development, and probably the cost, one should seek to acquire the rights to operate a system already tried and tested. In selecting a system ease of operation and adaptation to national circumstances (e.g. the language used by the operators as well as the national classifications and their coding indexes) should be given priority. A (C)AC system’s operational advantage is likely to depend more on the type and cost of data registration and the quality of the coding indexes than on any particular feature of the search and decision making algorithms of the system. Fast response times and an easy-to-understand interface between the computer and the operator should also be important considerations when selecting a (C)AC system.

5 The development and use of coding indexes

39. This section is based mainly on experience from English speaking industrialized countries, because the limited documentation readily available on the development and use of coding indexes and coding procedures have mainly originated in such countries. Our knowledge of the experience with coding in other languages and settings is limited, and it is therefore difficult to say to what extent the documented experience from English is transferable to other languages and

cultures. However, the suggestions below may provide a starting point for work and experiments with the coding of responses in other languages.

5.1 What is a coding index?

40. The process of coding the ‘industry’ and ‘occupation’ responses provided on the census questionnaires involves the task of matching the responses against the entries in the respective coding indexes, to find the appropriate codes. Thus the coding indexes are the key instruments for this matching process. The indexes can take the physical form of durable printed publications, loose-leaf binders, computer print-outs or machine-readable files within a computer system, and the matching can be carried out by a person, i.e. the coder, by a computer or through interaction between the coder and a computer.

41. Most detailed ‘industry’ and ‘occupation’ coding is still carried out using clerical procedures: Relevant information is written on the questionnaire by respondents or enumerators/interviewers and brought to one or more central offices for coding. Clerical staff (coders) scrutinize each case, decide (suitably guided by a coding index and instructions) to which ‘industry’ or ‘occupation’ group to allocate it and record the appropriate code on a document, or directly on to a computer-readable medium, for further processing.

42. The coding index is the principal instrument for linking the words used in the various parts of a response to the numerical code which represents the allocation of that response to the corresponding group of the classification. The coding index guides the coder by listing information, e.g. key words, which can be found in the responses. It indicates how different responses are allocated to the detailed or more aggregate groups of the classification, depending on the nature of the information in the response and on the instructions for the coding process. Thus a census coding index must be a reflection of the type of responses which one will find written on the census form by the respondent, or written by the enumerator on the basis of the information received from the respondent. **The ‘industry’ (‘occupation’) coding index should reflect the type of words and expressions which the respondents will use when asked to give the information about their place of work (or their job) which the census questions ask for.**

43. It is important to recognize that a coding index is different from the list of groups specified in the classification. The titles chosen for those groups are designed to be as descriptive as possible for the group content, given that only a few words may be used. Only a few of these titles will correspond to the terms used by individuals when asked about the activities of their place of work or about the main tasks and duties of their job. The coding index is also different from a list of titles or terms which may have been chosen to illustrate the content of the groups. This list may also contain entries which are never used as a title by any person describing their job or place

of work. Nevertheless, this type of list may serve as a useful starting point for the construction of a coding index.

44. Since they have to be in place before the census coding operations start, the coding indexes have to be constructed in anticipation of what the census responses will look like. Their basis therefore will have to be actual responses to similar questions in the last census, in household surveys carried out after that time or in census pre-tests. It will also be useful to collect terms and expressions concerning types of economic activities and jobs which can be found in advertisements for products and services (for the 'industry' index) and vacancies (for the 'occupation' index), registrations of vacancies and job seekers in employment offices, etc.. The collection and correct coding of the elements to be included in the coding indexes has to be carried out by experts on the respective classifications. The work will be painstaking and time consuming, but the investment involved in the collection and coding of up to 10, 20 or even 30 thousand index entries ahead of the census will prove well worth the effort in terms of the speed and reliability with which hundreds of thousands or millions of census forms can be coded during census operations.

45. It is in many ways correct to see the coding index as the ultimate manifestation or embodiment of a classification. However, it should be regarded as a working tool and not be given the status of being an official part of the classification, for two reasons: (i) It will be necessary to update the coding index during the census operations, and when it is used later in other coding operations. This flexibility may be difficult to achieve if it is a formal part of the classification. (ii) In order to reflect actual responses a coding index may need to include terms, e.g. brand names, which are commonly used to describe places of work or types of jobs, but which may be protected as trade marks, by copyright or for other reasons be difficult to include in an official publication. Such problems will usually not arise when the coding index is used only as an internal working document, and is made available to other organizations only for use as such.

46. At the start of the census operations the coding indexes must be assumed to be incomplete, and provisions must be made to up-date them during the whole period of the census coding operation, and most frequently and with more new items early in the process. The updating should be an extension of the query resolution process in the sense that the nature and outcome of resolved queries should be made available to all coders as soon as possible, in case they encounter the same type of response. The best is to issue new versions of the coding indexes frequently. Note that to issue a complete, new version of a coding index is better than to issue additions to the index, because the new entries will belong in different places in the index and the coders will have to transfer the information from the note on additions into the main coding index, with the danger that they will be making mistakes. New, complete versions of the coding indexes issued frequently in the first weeks of the coding operation will also reduce the danger that individual coders will keep their own notes on the coding of particular responses. Such notes can easily be the source for systematic differences between coders in coding responses not reflected in the initial version of the coding indexes.

47. During the census operations the physical form of the coding indexes must reflect the temporary nature of the version currently in use at any point in time: Each issue should be precisely dated and when choosing the form of the paper versions the main consideration should be the speed of reproduction. Both paper and electronic versions should be issued with orders to destroy earlier versions. Only when it is clear that further significant additions to the coding indexes will not be forthcoming during the remainder of the census coding operation should they be issued in a form more suited for dissemination to other users, such as various government departments, survey agencies and academic users. Then a well-bound quality publication may be considered, but a format more compatible with their status as working tools, and with regular, although less frequent, updating may be preferred, e.g. ring-files or computer print-outs may be the appropriate format for paper versions.

5.2 Developing and updating the ‘occupation’ coding index

48. The process of developing and updating a coding index for a classification of occupations should be viewed as one part of the general process to maintain and update the classification. As new technologies are introduced, or new ways of organizing work between or within enterprises, new types of jobs will appear with new combinations of tasks or new types of tasks. Jobs belonging to these new types may be given new titles by their incumbents or their employers, or may be referred to under existing job titles. At the same time existing jobs may be given a new title without their tasks and duties having changed in any significant manner, e.g. as a result of reorganization of the enterprise or because the jobs’ placement in a wage hierarchy or a collective agreement has been changed. Thus, there is a need to keep track of job titles and the associated job descriptions, and to monitor the relationship between this information, the entries in the coding index and the associated occupational groups and codes. Unfortunately, often this work has not been undertaken by the custodian of the national standard classification of occupations, nor by anyone else. This will mean that not only the updating but also the creation of the occupation coding index may have to be undertaken from scratch for the census.

49. A full scale job content mapping exercise cannot be undertaken as part of the census preparations. That would be too time-consuming and costly. The most realistic alternatives will be to carry out (i) post-coding reviews of recent survey operations; (ii) reviews of job vacancy notices; and (iii) consultations with job placement services:

50. Post-coding reviews make use of the tools and members of the coding team(s) used for recent survey(s). Coders are a good source of information on the adequacy of the coding index and other tools which they use. Their suggestions for improving the index and for additional or revised entries should be recorded and investigated. Ideally preparations for a post-coding review should have been designed into the data processing operation of the survey in question. It is essential that information which may be useful in updating the classification and coding index be

carefully collated and retained during survey operations. Useful information can be found in the records of problems encountered, queries raised and decisions and amendments to the working instructions adopted in the course of coding. Procedures for capturing this information should be made part of the normal routine for continuous and regular surveys, e.g. Labour Force Surveys, as well as for the registration of occupations which takes place in the local offices of the employment services.

51. When available from job advertisements in newspapers, journals, bulletin boards and/or registrations at a local office of the employment services, job vacancy descriptions may provide a useful source for constructing or up-dating information on job titles and detailed job descriptions, and therefore also a coding index. This is especially the case when the notices have been coded to the occupational classification as part of the job vacancy recording process in an office of the employment service. From such notices about job vacancies it should be possible to find a reflection of the impact of technical and organizational change on the allocation of tasks to jobs, and to develop proposals for new entries for the index (and the classification). One advantage of this approach is that it does not require expensive initial search for contacts, as follow-up inquiries to a vacancy of interest can use the name, address and contact person of the employer found in the vacancy notice. The main disadvantage is that job vacancies which have been advertised in newspapers or on bulletin boards, or notified to and recorded at an employment agency, normally only cover a limited range of occupations and industries.

52. In some countries the employment agencies have established standard procedures for the collection of job vacancy material. For example, when employers contact an employment agency, the agency may create a computerized record of information containing the job title and a brief description of the main duties or tasks associated with that job title. These records may then be used within a word-processing system or may form part of a database of vacancy information. Those responsible for creating the occupation coding index for the census operation should therefore investigate whether relevant material can be found in the employment services as well as elsewhere.

53. When organizing the material to be included in the coding index of occupations the first issue concerns the structure of the index itself. Basically the choice is between two different approaches: In some statistical agencies the approach has been that the index should be all-inclusive: i.e. every distinct type of response found in the process of coding should, in theory, have an entry in the index. (Exceptions may be made for misspellings and inversions of words which are without consequence for the meaning of the response.) An advantage of this approach is that it may be possible for coders to find that title and/or those tasks listed in the index, even when faced with an obscure job title and/or task description. The main disadvantage is that the size of the index may become very large and its sheer size may slow down the process of searching for the "right" entry in the index, and thereby slow down the coding, whether the coding is done manually or is computer assisted. Also, large verbatim indexes may create the impression that coding is a simple task, involving a straightforward matching between a response and an index

entry. However, no matter how large the index (and some have been developed which contain over 30,000 entries) it will always be the case that a significant proportion of responses fail to match the index entries exactly, and for those one has to use rules and/or judgement to make the 'best' match. For these reasons, in other agencies the approach has been to develop a structured index.

54. A structured index does not try to reflect every possible response, but the entries are accompanied with instructions to the coder on how to break down the available response into functional (key) words and qualifying nouns or adjectives for effective search for the most appropriate code. The primary entries in the index are the functional words. If a functional word in itself is not sufficient to uniquely identify the group, an appropriate qualifying word (or phrase) must be added to distinguish between the possible alternatives having the same functional word. If this is not sufficient to resolve all ambiguities, second or higher order qualifying words should be used. The following examples may illustrate the system for transforming an occupational response into an entry in a structured coding index according to the following format:

Response: Y Functional word/1st qualifying word/2nd qualifying word:

Examples:

Cost accountant: Y accountant/cost
Drilling machine operator: Y operator/machine/drilling
Aircraft instrument maker: Y maker/instrument/aircraft
Room maid: Y maid/room
Marine biologist: Y biologist/marine
Capstan lathe setter-operator: Y setter-operator/capstan lathe

55. The following examples from the coding index used for coding occupation in the 1986 Population Census in Australia will illustrate the use of the qualifying words as well as the way instructions about the use of the index can be incorporated with the index entries (the codes given are those of the *Australian Standard Classification of Occupations, first edition - ASCO)*:

5999	Researcher/market/interviewing
2909	Researcher/market/statistician
2907	Researcher/market (except above)

(The second of these entries could also be listed as: 2909 Statistician/market/research)

2701	Researcher/accountancy
2107	Researcher/agricultural
2907	Researcher/anthropology
2999	Researcher/assistant to parliamentarian
2107	Researcher/biological sciences (except medical)
2101	Researcher/chemistry (except medical)

2109	Researcher/medical
3103	Researcher/toxicology
2000	Researcher (no additional information about type of research)
8919	Restaurateur/assisting in kitchen
4705	Restaurateur/cooking
1503	Restaurateur/supervising staff and administration
6505	Restaurateur/waiting on tables
1503	Restaurateur (no additional information about specific tasks)
3999	Retoucher/photographic
4503	Retoucher/printing
1311	Secretary/assistant/senior govt. officer/computing div.
1307	Secretary/assistant/senior govt. officer/distribution div.
1313	Secretary/assistant/senior govt. officer (except above)
6503	Secretary/club/tending bar
1599	Secretary/club (except above)
1201	Secretary/trade union
5601	Secretary/receptionist
5101	Secretary (no additional information about specific tasks)
5101	Secretary (except above)
4405	Sign writer
4921	Silverer/glass
4923	Silversmith
2815	Singer.

56. The structured coding index will normally have a much smaller number of index entries than a complete listing index. This is the result both of the restriction of the index to functional words when possible and the use of '(except above)' instructions. They allow the exclusion from the index of a large number of responses where the qualifying words are immaterial for the selection of the correct group.

57. In a structured index for coding occupation the functional word is the word in the relevant response which alone can serve as an occupational title, however imprecise. The qualifying words

usually will indicate some form of specialization or tasks. Sometimes the functional word may be precise and in itself suffice as an index entry, cf. 'Sign writer' in the examples above. However, the functional word may also be very ambiguous, cf. the “Researcher” and 'Secretary' examples above. Note that the qualifying words in some of the 'Secretary' examples do not serve to distinguish between specializations, but between occupations which are very different in nature.

58. The construction of the structured coding index must reflect and support the coding rules to be used for assigning the occupation codes on the basis of the responses to the relevant questions, and use as qualifying words permissible ancillary information given in other responses and indicated in the index. This means that one should organize the index alphabetically first in terms of functional words, then in terms of the first qualifying word with those entries which also have a second qualifying word listed before those which do not, and the'(except above)' instruction should follow the entries with qualifying words. In English it will be normal to use as functional words that can be found in either (as first priority) the title component of the occupational response, or (second priority) the task component. First priority for qualifying words will be those normally found in the task components of the response. Second priority for qualifying words should be given to words found in the ‘industry’ or ‘name/type of employer’ response, and then with due regard to the rules established for the use of such information when coding ‘occupation’.

59. The advantages of having a structured coding index are twofold: First, it helps the coder to search for index entries in a way which is consistent with the coding rules. Second, it speeds up the task of coding by restricting the number of entries the coder has to search through in the index, because of the smaller number of entries.

60. Some words may be found to be in common usage in job titles, but can be ignored for the purpose of creating a coding index. For example, words such as 'boy', 'girl’, 'man', 'woman', 'worker' and 'executive' do not carry information about the tasks which constitute a particular job. It is usual to exclude such words from the index and to instruct coders to ignore them.

5.3 Developing and updating the ‘industry’ coding index

61. Most census coding operations will find it useful to have two “coding indexes” for the coding of industry:

(a) a list of as many as possible of the establishments which, at the time of the census, are/were operational in the geographic region covered by the coding operation, where each establishment has been given the correct industry code by those who are specialists in establishment surveys and the coding of establishments’ activity. In practice such lists will often cover only large, formal sector establishments as they have to be created from information kept in tax offices, licensing offices, chambers of industry and commerce, etc. They may nevertheless cover significant proportions of the work force, and their use for census coding will eliminate one possible source of inconsistency in employment statistics between the census results and the results of establishment surveys.

(b) a list of significant word combinations reflecting the answers given and written down by enumerators to 'industry' questions on the census form, such as those presented in section 7 of *Gilbert (2001)*. This will be an index of the same type as that created for the coding of 'occupation', described in section 3.5.2 above.

62. In the same way as for 'occupation' the process of updating the coding indexes for 'industry' responses should be viewed as part of the general processes required to maintain the 'industry' classification. As new technologies are introduced as well as new ways of organizing work between or within enterprises, new products and new services will appear and be provided by separately identifiable units, i.e. separate establishments. New establishments will be created and old ones will go out of business. Establishments may change their products while continuing to operate under existing names, or may be given a new name without their function, products or services having changed in any significant manner, e.g. as a result of reorganization of the enterprise or change of ownership. Thus, there is a need to keep track of establishment names as well as activity descriptions and associated designations, and to monitor the relationship between this information, entries in the coding index and the associated industry groups and codes. The work to maintain establishment lists of type (a) above are normally carried out by those responsible for an establishment register or for establishment surveys. However, the corresponding work with respect to lists of type (b) has often not been done, neither by the custodian of the national standard 'industry' classification, nor by anyone else. This means that not only the updating but also the creation of this type of 'industry' coding index must be undertaken from scratch for the census.

63. A full scale establishment mapping exercise cannot be part of the census preparations. This would be too time-consuming and costly. The most realistic alternatives for updating (b) will be to carry out (i) post-coding reviews of recent household survey operations which included an open industry question; and (ii) reviews of advertisements and notices in newspapers and other media. For updating lists of type (a) it will be necessary to extract as much help and information as possible from those responsible for establishment surveys and registers in the statistical services. Depending on national circumstances it may also be useful to seek the help of national and local tax authorities as well as chambers of commerce, etc.

64. Notices and advertisements in newspapers, journals and/or bulletin boards about their products and services, creation and expansion, or about vacancies, will be used by many formal sector businesses. Such sources may therefore be useful for constructing or up-dating information on establishments and their activities. Unfortunately, information about small informal sector establishments cannot be found there.

65. To be useful in the coding process the list of establishments should for each unit give both a name and the physical location. Location should be indicated by a street address if possible, or by naming the (smallest) district in which the unit is located or is operating. If alternative forms of the name, such as abbreviations, initials, old names etc are used or recently have been in use, they should also be included in the list as separate entries, because of the possibility that persons working for them may use these variants in their answers. The entries in the list should be organized alphabetically, with clear rules for where to find entries consisting of initials and abbreviations.

66. Post-coding reviews make use of the tools and members of the coding team(s) used for recent survey(s). Coders are a good source of information on the adequacy of the index and other tools which they have used. Their suggestions for improving the coding index and for additional or revised entries must be recorded and investigated. Preparations for the use of post-coding review procedures for the development of coding index material should be designed into the data processing operation of the surveys. The useful information typically consists of records of problems encountered, queries raised and decisions and amendments to the working instructions adopted in the course of coding. The procedures to capture this information should be part of the normal routine for continuous or regular surveys, both establishment surveys and the Labour Force Surveys, as well as for the registration of establishments and activities which takes place in the local tax offices, licensing offices or chambers of commerce.

67. When organizing the material to be included in the type (b) coding index of activities (‘industries’) the first issue concerns the structure of the index itself. Basically the choice is between two different approaches: One approach is to make the index all-inclusive: i.e. every distinct type of response found in the process of coding should, in theory, have an entry in the index. (Exceptions may be made for misspellings and inversions of words which are without consequence for the interpretation of the response.) One advantage of this approach is that it may be possible for coders, when faced with an obscure type of activity or product, to find those terms listed in the index. The main disadvantage is that the size of the index may become very large and its sheer size may slow down the process of searching for the "right" entry in the index, and thereby slow down the coding, whether the coding is done manually or is computer-assisted. Also, large indexes create the impression that coding is a simple task, involving a straightforward matching between a response and an index entry. However, no matter how large the index it will always be the case that a significant proportion of responses fail to match the index entries exactly, and one has to use rules and/or judgement to make the 'best' match. For these reasons, an alternative approach may be to develop a structured index.

68. A structured index does not try to reflect every possible response, but the entries are accompanied with instructions to the coder on how to break down the available response into functional (key) words and qualifying nouns or adjectives. The primary entries in the index are the functional words. If a functional word in itself is not sufficient to uniquely identify the industry group, an appropriate qualifying word (or phrase) must be added to distinguish between the

possible alternatives having the same functional word. If this is not sufficient to resolve all ambiguities, second or higher order qualifying words should be used. The following examples may illustrate the system for transforming an ‘industry’ (or ‘type of activity’) response into an entry in a structured coding index according to the following format:

Response:Y Functional word/1st qualifying word/2nd qualifying word:

<u>Examples</u>:

Sheep farm:Y sheep/farm
Car rental agency:Y car/rental
Youth club:Y club/youth
Tax assessment office:Y office/tax/assessment
Cleaning service:Y cleaning/services
Cleaning products production:Y cleaning/products/production

69. The following examples have been taken from the coding index used for coding industry to the *UK Standard Industrial Classification of Industrial Activities 1992.*:

15.11/1	Abattoir
74.40	Advertising/agency
74.40	Advertising/agent
74.40	Advertising/campaigns/creation
74.40	Advertising/campaigns/realization
74.40	Advertising/consultant
74.40	Advertising/contractor
92.11	Advertising/film/production
31.50	Advertising/lights/manufacturing
74.40	Advertising/material/design
22.22	Advertising/material/printing
22.22	Advertising/newspaper/printing
22.12	Advertising/newspaper/publishing
74.40	Advertising (no further information)

70. In a structured index for coding ‘industry’ the functional word is the word in the relevant response which alone can serve as a designation of a service, a product or a function, however imprecise. The qualifying words usually will indicate a special form or variety of a product or service and/or a type of activity associated with the product or service. This sequence of words has been chosen because the number of different designations for activities is much smaller than the number of designations of different products, services and functions. Sometimes the functional word may be precise and in itself suffice as an index entry, cf. 'Abattoir' in the example above. However, the functional word may also be very ambiguous, cf. the “Advertising” examples above.

71. The construction of the structured coding index must reflect and support the coding rules to be used for assigning the ‘industry’ codes on the basis of the responses to the relevant questions. When necessary, permissible ancillary information given in other responses can be used if indicated in the index as qualifying words. This means that one should organize the index alphabetically first in terms of functional words, then in terms of the first qualifying word with those entries which also have a second qualifying word listed before those which do not, and the '(except above)' instruction should follow the entries with qualifying words. The functional words listed in the index must reflect those which can be selected from the “type of activity” parts of the responses, and the qualifying words must reflect those which can permissibly be selected from the response to questions about main products or functions at the place of work.

72. The advantages of having a structured coding index are twofold. First, it helps the coder to search for index entries in a way which is consistent with the coding rules. Second, it speeds up the task of coding because of the smaller number of entries in the index.

5.4 Using the coding indexes

73. Coding can be seen as a process where the task of the coder is to 'translate' the information provided by the recorded responses to the most appropriate code in the relevant classification structure. The main tools for this 'translation' are the coding indexes and the coding instructions - including instructions on when a response should be treated as a query to be resolved by supervisors or expert staff. The instructions should specify:

(a) how this translation process should be carried out;
(b) what items to look for in the relevant response and in what order; what type of ancillary information to use from other responses;
(c) when such ancillary information can permissibly be used and how to use it.

Ideally the coding index should have been constructed to reflect and support the use of these instructions, as described in the previous sections on the development and updating of the ‘occupation’ and ‘industry’ coding indexes.

5.4.1 Using the ‘occupation’ response when coding ‘occupation’

74. The starting point should always be the response given to the ‘occupation’ question(s), i.e. question(s) about “what kind of work” the job involves and the usual or main tasks and duties carried out in the job. This should normally result in a response consisting of a job title and a few words on main tasks. When using an “all inclusive” coding index the coder should start by marking those words which are relevant for the search in the index. When using a “structured”

index the coder should start by identifying the “functional word” part of the response. This will normally be a job title, or a word which easily can be converted to a job title. Then the coder should look for this word in the index. The tasks part of the response should either be used to supplement or modify the information provided by the title, or be transformed to a title, e.g.. 'baking bread' to 'baker, bread', 'cleaned school' to 'cleaner, school'. Transformation of a task response to a title should be performed when there is no proper title response or when there is no index entry corresponding to the title given, as may be the case of 'labourer', 'civil servant', 'helper' and other non-informational titles. If the occupational responses are not sufficient to determine a detailed occupational group then the coder should choose one of three alternatives:

(a) For further clarification look at the form for recorded ancillary information of a specified type.
(b) Use an appropriate code for inadequately specified responses.
(c) Refer the case to supervisors as a query.

The coders should be given clear instructions on the proper alternative to choose in the different situations.

5.4.2 Using ancillary information on ‘industry’ or name and type of employer when coding ‘occupation’

75. Most modern occupational classifications, including ISCO-88, are designed on the principle that 'occupation', meaning a particular pattern of work tasks and duties which constitute an individual's job, should be kept conceptually separate from 'industry', meaning the type of economic activity to which the job contributes. Thus, an 'electrical maintenance fitter' may work in any of a range of different industries and this person's occupation cannot be validly deduced from a knowledge of the industrial category of the employing organization. Without breaching this principle, it must nevertheless be recognized that certain occupations are to be found solely or predominantly in particular industrial sectors. In such cases knowledge of the industry may clarify an inadequate occupational title or description for coding purposes. For example a 'face worker' working in a coal mine may be deduced to be a miner engaged in coal cutting. In other cases the descriptions of work activities used to identify an occupation are best formulated in terms of, for example, the nature of the material worked with (e.g. wood, rubber, leather etc.). This information may be deducible from a knowledge of the industrial sector in which the job is located and again help to clarify a vague occupational response. For example, the occupational term 'coil winder', used on its own, is ambiguous because coding depends on whether the wounding is some form of metal wire, some form of textile product, etc. Knowledge that the job is located in the a textile manufacturing establishment may be sufficient to resolve the ambiguity with a reasonable degree of certainty.

76. Because some interrelationships between ‘occupation’ and ‘industry’ are inherent in the industrial structure of the economy, they can be made use of to improve the accuracy of occupational coding. However, there are costs as well as benefits in this practice: In the first place, there is always some danger that inferences from ‘industry’ to ‘occupation’ will be based on incorrect or out-of-date assumptions about the distribution of occupations across industries. In the second place, when coding is being done on a large scale, coding work rates and inter-coder consistency are important considerations. Coding rates are likely to be slowed if the coder routinely will have to consider extra sources of information on the form in order to arrive at a code, particularly if the extra information is itself hard to interpret. In such circumstances there is also a danger of increasing inter-coder variability, since different coders will tend to interpret the ancillary information in different ways. These latter two problems are minimized:

(a) if industry is coded in advance of or at the same time as occupation, so that no further interpretation of the industry responses is required for the occupation code;

(b) if coders are allowed to use data on ‘industry’ only where the responses to the specific questions on job title and activities are inadequate to determine an occupational code; and

(c) if the choices that can be made when coding ‘occupation’ on the basis of the ‘industry’ code are exhaustively specified through index-referenced instructions.

77. A simplified example may clarify this: A coder encountering the job title 'coil winder' would be instructed in the coding index under 'winder, coil' to look first at the description of job activities for information on the type of material wound. In cases where no indication of this was found the coder would look next at the pre-allocated industry code. If this was code 'x', standing for 'textile manufacturing', the occupational category would be determined as 'textile yarn winder'; if the industry code was 'y', standing for 'electrical machinery manufacturing', the occupational category would be determined as 'wire winder'. If the industry code was other than 'x' or 'y' the occupation would be placed in an appropriate 'inadequately described' category by the coding index.

5.4.3 Use of other ancillary information when coding ‘occupation’

78. Some coding operations include information about the educational and vocational qualifications of respondents among the ancillary information which it is permissible for coders to use to determine the appropriate occupational code. Again, this should be based on detailed knowledge of the relationships between training and qualifications on the one side and the corresponding occupations on the other. In all countries this relationship varies between occupations, and in most countries the relationship is close only for a limited number of occupations. Even when the relationship is close it must be recognized that the fact that a person has a particular qualification does not mean that his/her job will include the corresponding tasks.

(A person with a medical degree who is working in a hospital may not have healing tasks. This may be because he/she has been promoted to a job which consists of management tasks, or it may be because he/she could not get a job corresponding to the type and level of technical training, e.g. because he/she lacks necessary language skills.) The use of information on qualifications or educational attainment as ancillary information should therefore be very carefully controlled and probably restricted to be used by expert coders in query resolution.

5.4.4 Inadequate responses and queries when coding ‘occupation’

79. Some responses simply cannot be coded to a detailed occupational group. This will normally be for one of the following reasons:

(a) the response may be vague, i.e. not contain enough information to be coded according to the coding index and coding rules; or
(b) the response may be precise, but may use a title and/or indicate types of tasks or combinations of tasks which do not correspond to any of the index entries.

80. Unfortunately, the number of cases of type (a) is likely to be quite substantial, even with well-formulated occupation questions and well-trained enumerators. In order to keep to manageable proportions the number of queries which the supervisors and expert coders must handle, the coding index and the coding instructions should be designed to guide the coders with respect to the most common of such cases. The simplest solution will be to specify that the response should be coded to a 'default' group, cf. the examples of 'Researcher', 'Restaurateur' and 'Secretary' in section 5.2. This default group may in some cases be a specific detailed group, because this reflects the dominant usage of the terms found in the response, cf. the use of '1503 Restaurateur' and '5101 Secretary' as default groups in those examples. However, the default group will often have to be one of the aggregate groups in the classification, because it is not possible to identify one particular detailed group as dominant within the aggregate group indicated. In the Australian example above the entry '2000 Researcher' indicates that a response giving only “researcher” as information could only be coded validly to ASCO major group 2. Similarly a response like 'Clerk, clerical work' would normally have to be coded to the aggregate group for 'Clerks', unless the industry response gives very clear information about the type of clerical work done (which is not likely).

81. There is a real danger that before they even try to find a precise code coders may use 'default' groups as 'dump groups' for responses which are difficult to code. Some coding operations therefore have tried to keep the coders ignorant of the possibility of using such codes and only allowed them to be used by better trained supervisors. However, this strategy may create a morale problem among the coders, and place a very large query burden on the supervisors. It may therefore be better to monitor carefully the use of 'default' codes by the coders.

Asking them to record the index item number selected as most valid for the response, rather than the code given by this item, may also restrain the temptation to “code by memory”.

82. The fully specified responses which are not adequately covered by the classification and the coding index should always be handled by expert coders. Their appearance should be recorded carefully, both to ensure consistent treatment of equal cases and because these cases represent an important source of information for the updating of the coding index as well as of the classification itself. These cases can either be handled by using the priority rules specified for the classification or by assigning them to an appropriate group for occupations ‘not adequately covered" by the classification. As indicated the latter solution should only be used by expert coders. It is important to note that these groups are not the same as the 'not elsewhere classified (nec)' groups of the classification. Great care must be taken not to make a confusion between the two types of groups, as the nec group for e.g. “minor group xyz” are intended to include only the precisely defined occupations belonging to that minor group, but which are too small to warrant separate unit groups within it.

83. Priority rules can be applied to some of the responses which indicate task combinations which cut across the groups defined in the classification, e.g.. 'baker, baking/ selling/managing shop'. Most classifications based on ISCO-88 will specify priority rules in terms of tasks performed for the allocation of such jobs to occupational groups. In ISCO-88 it is specified that priority should first be given to the tasks which require the highest skill level, and secondly that production-oriented tasks should be given priority over managerial or administrative tasks. 'Main tasks', in terms of e.g. time spent, are not to be given priority unless they completely dominate, both because an employer is likely to be concerned that a worker can carry out the most skilled tasks required, even if they are only seldom activated, as in emergencies, and because time allocation of tasks is an information normally not available. Thus in the above example the code to be specified in the coding index should be the code for “baker”.

5.4.5 Using the industry response when coding industry

84. The starting point for coding industry should always be the response given to the first part of industry question(s), i.e. question(s) asking for the name and geographical location, e.g. street address, of the place of work. If the name list provides an exact match on both name and location, then the industry code given to this unit in the name list can be given to the response. If there is no exact match then the coder should make use of the regular coding list for industry, by selecting a word from the response which provides information about the type of products, services or function which the unit produces or provides. If this is not sufficient to determine a code, as with “advertising” in the example above, then the coder should identify supplementary words, qualifiers, which may give more precise information about the product, and/or the type of process involved. If the industry response does not contain information sufficient to determine a

detailed industry group then the coder should choose one of three alternatives:

(a) Look at the form for recorded ancillary information of a specified type for further clarification.
(b) Use an appropriate code for inadequately specified responses.
(c) Refer the case to supervisors as a query.

The coders should be given clear instructions on the proper alternative to choose.

5.4.6 Using information on ‘occupation’ when coding ‘industry’

85. Most modern industry classifications, including ISIC, rev.3, are designed on the principle that 'industry', meaning a particular set of productive activities resulting in one or more products or services produced by an economic unit, should be kept conceptually separate from 'occupation', meaning the type of work performed by a person working in the establishment. Since many different “occupations” may be represented in the same establishment, one cannot validly conclude, normally, from one person’s occupation to the industry of the workplace, even if this happens to be an occupation which tend to cluster in a particular industry, e.g. bus drivers. However, there are a few exceptions to this, e.g “university teachers” are only found in the “education” industry, ‘police officer’ only in “Public administration and police”, and taxi drivers only in transportation. For some “own-account” workers there will be a direct link to a particular industry, e.g. own account plumbers should logically only work in the construction industry. Such cases can be identified and incorporated into the industry coding index, through the use of qualifying words and the rules for its use.

86. In this way some interrelationships between occupation and industry, which are inherent in the industrial structure of the economy, can be made use of to improve the industry coding. However, there are costs as well as benefits in this practice: In the first place, there is always some danger that inferences from occupation to industry will be based on incorrect or out-of-date assumptions about the uniqueness of an occupation to a particular industry. In the second place, when coding is being done on a large scale, coding work rates and inter-coder consistency are important considerations. Coding rates are likely to be slowed if the coder routinely will need to consider other responses in order to arrive at a code, particularly if those responses themselves are hard to interpret. In such circumstances there is also a danger of increasing inter-coder variability, since different coders will tend to interpret the information in differently. These latter two problems are minimized:

(a) if occupation is coded in advance of or at the same time as industry, so that no further interpretation of occupation responses is required for the industry code;
(b) if coders are allowed to use information on occupation only where the responses to the

specific questions on job title and activities are inadequate to determine an industry code; and

(c) if the choices that an industry coder can make on the basis of the occupation information are exhaustively specified through index-referenced instructions.

87. A simplified example may clarify this: A coder encountering the establishment name 'Institute for marketing studies' might be instructed in the coding index under 'Institute/marketing' to choose between the code for “marketing” and the code for “education”. If the occupation is given as “manager”, “secretary” or “janitor” there would be no basis in the occupation information to make the choice, as all three types of occupations may exist in either industry. However, if the occupation is given as “accounts executive” then it is likely that the establishment is a marketing firm (with a fancy name), and if the occupation is given as “professor” or “lecturer” then it is likely that the establishment is a training institution.

88. Further possibilities for using “occupation” as ancillary information for industry coding may exist in “one company” locations where it may be possible to have the range of occupations which will only be found in the dominant company’s establishments. This situation will reduce or eliminate the possibility that e.g a bus driver, living in this location, could be employed by any other type of establishment than the local bus company, as this company runs all busses in the area.

89. No other information which may be available on the census questionnaire seems relevant as ancillary information for the coding of industry.

5.4.7 Inadequate responses and queries when coding ‘industry’

90. Some responses simply cannot be coded to a detailed industry group. This will normally be for one of the following reasons:

(a) the response may be vague, i.e. not contain enough information to be coded according to the coding index and coding rules; or

(b) the response may be precise, but may indicate types or combinations of products, services or functions which do not correspond to any of the index entries.

91. Unfortunately, the number of cases of type (a) is likely to be quite substantial, even with well-formulated industry questions and well-trained enumerators. In order to keep to manageable proportions the number of queries which the supervisors and expert coders must handle, the coding index and the coding instructions should be designed to guide the coders with respect to the most common of such cases. The simplest solution will be to specify that the response should be coded to a 'default' group. This default group may in some cases be a specific detailed group

because this reflects the dominant usage of the terms found in the response, cf. '74.40 Advertising' as default group in the example above. However, the default group will often have to be one of the aggregate groups in the classification, because it is not possible to identify one particular detailed group as dominant within the relevant aggregate group.

92. As with ‘occupation’ there is a real danger that 'default' groups may be used by coders as 'dump groups' for difficult to code responses before they have tried to find a precise code. Some countries have therefore tried to keep the coders ignorant of the possibility of using such codes and only allowed them to be used by better trained supervisors. This strategy may create a morale problem among the coders and a very large query burden on the supervisors. It may therefore be better to monitor carefully the use of 'default' codes by the coders.

93. The precise responses which are not adequately covered by the classification, i.e. cases of type (b) above, should always be handled by expert coders and recorded carefully, both to ensure consistent treatment of equal cases and because these cases represent an important source of information for the updating of the coding index as well as the classification itself. During the coding operation these cases can either be handled by using the priority rules specified for the classification or by assigning them to one or several groups for activities ‘not adequately covered" by the classification. It is important to note that these groups are not the same as the 'not elsewhere classified' groups of the classification, which include clearly defined activities which are not important enough in scale to be given a separate identity within the larger group to which they belong. Great care must be taken not to confuse the two types of groups.

94. Priority rules are difficult to apply to the responses which indicate that the place of work produces combinations of products and/or services which cut across the industry groups defined in the classification, e.g. 'repairing cars & selling petrol'. Most industry classifications based on ISIC, rev.3 will say that such cases should be resolved with the help of priority rules formulated in terms of contribution to value added of the enterprise, or the number of persons employed in the different activities. However, this information will not be available to the coders or their supervisors. One solution may therefore be to refer these cases to the classification specialists and they may be able to identify the establishment and on that basis determine a code.

5.5 Level of coding

95. The more information retained after the coding, the more valid and therefore valuable can the resulting statistics be for the users. **The coding process should therefore be designed to find and record the most detailed codes supported by the responses**, even though traditionally the most common procedure has been to decide that coding should be done at a particular level of the classification structure, e.g.. the 3 digit level, no matter what information has been provided in a response. The arguments for this have commonly been: (i) that it would be

too costly to code to a larger number of groups, both in terms of coding errors and in terms of staff hours required; (ii) that the responses would not support coding to more detailed groups; and (iii) that (when coding only a sample) it would not be possible to publish results for the more detailed groups because of lack of observations. However, closer examination of these arguments in the light of the experience gained by statistical agencies has shown that:

(a) The marginal costs of coding to a larger number of groups in the classification, i.e. to a lower level of aggregation, are rather small in terms of increased error rate as well as in terms of work hours needed for coding and other costs, especially as measured against the increased validity and usefulness of the data. The error rate for aggregate groups does not seem to increase, on the contrary. One statistical agency has estimated that to code at the occupation level instead of the unit group level, i.e. to increase the number of possible groups from 280 to 1100, would require an increase in the coding index of only about five percent, with corresponding small effects on the cost of searching the index for the main detailed categories.

(b) Experience clearly shows that the ‘industry’ and ‘occupation’ responses found on a census questionnaire are very uneven in the amount of relevant information they will provide. Many responses will support detailed coding, especially if the questions are formulated along the lines presented in sections 5B and 7B of *Gilbert (2001)*. A significant number of responses will, however, not even support the level conventionally chosen. By insisting on a predefined level, the coding process may therefore both lead to unnecessary loss of information for a large proportion of the returns and to misrepresentation of the data quality for other returns.

(c) The similarity criteria used to define the groups in an ‘industry’ or ‘occupation’ classification are mostly defined with reference to respectively the nature of the production process and the type of work performed. Normally not much regard is paid to the number of employed persons in the resulting groups, except in the few national classifications where 'statistical balance' has been an important criterion when constructing the classification structure. This means that the number of jobs which can be found in groups defined at the same level in the classification, will differ greatly. The number of jobs in a group defined at an aggregate level may therefore be smaller than that of a group defined at a lower level in the structure of the classification, but within another aggregate group. In addition, the tabulation of ‘industry’ and ‘occupation’ statistics also typically involves both the merging of groups and cross-classification with other variables such as age, sex, region. Consequently one should not restrict tabulation possibilities during the coding process.

5.6 Use of nec groups and the coding of responses which coders cannot assign to

specified groups

96. Most industry and occupation classifications specify residual groups of “type x industry/occupations not elsewhere classified". As mentioned above these “nec” groups are designed to take care of activities and jobs that belong to the more aggregate groups, but which are not similar enough to any of the specified sub-groups within the aggregate group to belong in either of them, and they themselves include too few cases to warrant separately specified groups. **This means that nec groups should not be used to code those responses which the coders cannot assign to any of the specified groups.** Such responses can be either: (i) too vague and imprecise to allow the coder determine which group the job belongs to; (ii) indicate that the establishment (job) in question produces a combination of goods or services (involves a set of tasks) which cuts across the distinctions made in the industry (occupation) classification; or (iii) represent a type of production or work not covered by the classification.

97. The proper way of handling such not easily coded responses will depend on the type of case:

(a) Vague and imprecise responses should be coded to the level in the aggregation structure supported by the information contained in them - they should not be forced into any particular detailed group where it is likely that only a small proportion of these jobs would fall if one had an adequate response. For example, in the Australian 1986 Census 15 percent of the jobs coded to the major group ‘clerks’ could not be coded to any of the more detailed groups within this major group. It would obviously represent a significant distortion of the results if they had all been placed in one particular more detailed group together with those jobs which properly belonged to that group.

(b) The classification of establishments (jobs) with an uncommon mix of goods and services (tasks and duties) should, as far as possible be made on the basis of the general priority rules of the classification. Such responses should preferably be treated as queries and left to expert coders or the classification experts. Least disruption of the coding process is often made if these responses are given a special code and the questionnaires put aside for later examination by the experts. This treatment should also be given to responses which seem to represent establishments (jobs) with products, services or functions (tasks and duties) not covered by the respective classification. The reporting of difficult cases is an important input to the process of updating, maintaining and possibly expanding and revising the relevant classification.

5.7 Coding of subsistence jobs by ‘industry’ and ‘occupation’

98. As explained in section 2A of *Gilbert (2001)* the scope of ‘production’ and therefore

‘employment’ includes a number of productive activities with results mainly for own use or subsistence. This may create a problem when coding ‘industry’ to ISIC, rev. 3 or to a classification based thereon, as there are no separate groups for such activities in ISIC, rev.3. As the results are for own use (subsistence) the household is the productive unit in which these activities take place, and therefore the household serves the role of ‘establishment’ and is the unit for which an ‘industry’ code should be given. The ‘subsistence activities’ which are undertaken by members of the household will frequently consist of activities which, if they had been carried out by specialized units, should be classified to quite different groups in an ‘industry’ classification, e.g. ‘agriculture’, ‘food processing’, ‘construction’. However, as these activities are carried out in and for the household they must all be seen as being carried out by the same production unit. Thus the correct group for the unit should be determined by the ‘most important’ type of activity undertaken. “Most important” is supposed to be determined by proportion of “value added”, “total gross output” or “total employment”, but as this information will not be available in a census coding situation, it seems most relevant to consider agricultural production to be the most important subsistance activity. When using ISIC, rev.3 as the classification this will mean that the correct code will be “0130 Growing of crops combined with farming of animals (mixed farming)”, because it is in principle irrelevant whether or not the production is for the purpose of sale or not, see paragraph 28 in the “Introduction” to *United Nations (1989)*. However, for many descriptive and analytical purposes one may want to distinguish the subsistence activities from those which are more market oriented, in which case it seems consistent with the ISIC, rev.3 principles to give the subsistence activities the ‘industry’ code 0131. (Where subsistence activities are mostly linked to fishing and other water based activities the most relevant ISIC, rev.3 group would seem to be “0500 Fishing, operation of fish hatcheries and fish farms; services incidental to fishing”, in which case the code 0501 could be used to identify separately the corresponding subsistence activities.)

99. Based on a reasoning corresponding to the one outlined above for the coding of ‘industry’, it would seem that the most appropriate occupation group for “jobs” in subsistence activities would be “6210 Subsistence agricultural and fishery workers” when using a classification based on ISCO-88, unless the information available clearly indicates that the work performed is sufficiently specialized to warrant a different occupation code.

5.8 Coding to more than one classification

100. The principles and strategies outlined in sections 5.2 and 5.3 on the development and updating of coding indexes for “occupation” and “industry” are independent of the particular ‘occupation’ and ‘industry’ classification to be used when coding these variables. Consequently, it will be perfectly possible to provide each index entry with the code of more than one ‘occupation’ and ‘industry’ classification respectively. To do so will often be very useful, as the users of statistics may be interested in making comparisons of the results from the current census with

those from a previous census or survey, which may have used a different classification, or with the results from censuses or surveys from other countries. Coding can be done easily to each of the classifications of interest, e.g. the past and current version of the national classification and to the corresponding international standard classification, by assigning the codes for these classifications to each entry in the coding index and making sure that it is the identifier for the index entry which is being registered by the coders and the coding process. The code assignment (classification) can then be done by the computer during the tabulation by specifying the classification which one would like to use.

5.9 The problem of different languages

101. In the discussion above no reference has been made to the problems encountered in countries where there is more than one language which people use in their daily life, although this problem is referred to in several sections of *Gilbert (2001)*. On the assumption that interviewers/enumerators will know the language of the respondent and therefore can write down correctly in that language the answers to the industry and occupation questions, the best solution will be to make sure that the coding indexes can reflect those answers as given in the original language. Separate coding indexes for each major language may be the best solution when the use of these languages are highly concentrated geographically and coding can be organized accordingly. Otherwise the best solution may be to create "multi-language" coding indexes, to allow the coder to find an index entry which corresponds to what was written down as the response. While these coding indexes may be more difficult to construct than single language ones, the latter will require that someone, usually the enumerator, translate from the response given to the language of the index when writing the response. The problem with this is that the correct translation of occupational terms will not only require good general knowledge of the two languages in question, but also knowledge of the particular area of work, in order to understand precisely how particular terms on activities, products, services and jobs are used in the local context. Very few persons will normally be able to satisfy this requirement over the whole range of work situations covered by a population census. A multi-language coding index will be larger than a single language one, but it need not be dramatically larger, because in many countries the terminology reflecting modern sector activities and jobs will be common to many languages, and it will mainly be the terminology reflecting traditional activities and jobs which will differ. Such activities and jobs are normally less varied than those in the modern sectors, and the number of different terms less.

Suggestions for further reading

Australian Bureau of Statistics (1991): *Australian Standard Classification of Occupations (ASCO) - Expert Coding System: Occupation Level, Version 5.0 (DOS) on Floppy Disk.* Canberra, Catalogue no. 1226.0

Australian Bureau of Statistics (1993): *Australian Standard Classification of Occupations (ASCO) - Manual Coding System: Occupation Level.* Canberra, Catalogue no.1227.0

Australian Bureau of Statistics (1997): *ASCO - Australian Standard Classification of Occupations. Second Edition and ASCO Coder (Windows) on CD-ROM.* Canberra, Catalogue no.1220.0.30.001

Campanelli, P., K. Thomson, N. Moon and T. Staples (1997): "The Quality of Occupational Coding in the United Kingdom." Chapter 19 of Lyberg et al (1997).

Elias, P.; Halstead, K. and Prandy, K (1993): *CASOC - Software for Computer Assisted Occupational Coding.* HMSO, London.

Elias, P. (1996): *Automatic coding of occupational information for the 2001 Census of Population: a feasibility study.* Institute of Employment Research, University of Warwick, Coventry.

Embury, B.L. (1991): *The ASCO EXPERT Coding System.* Mimeographed Paper. Australian Bureau of Statistics. Canberra.

Embury, B.L. (1997): *Constructing a Map of the World of Work: How to Develop the Structure and Contents of a National Standard Classification of Occupations. STAT* Working Papers 2/95. International Labour Office, Geneva.

Gilbert, R. (2001): *Asking questions on economic characteristics in a population census*. STAT Working Paper 2001-1. International Labour Office. Geneva.

Hoffmann, E. (1994): “Mapping the World of Work: An International Review of the Use and Gathering of Occupational Information”, in Chernyshev, I., ed.: *Labour Statistics for a Market Economy: Challenges and Solutions in the Transition Countries of Central and Eastern Europe and the Former Soviet Union.* Central European University Press. Budapest, New York.

Hoffmann, E. et al (1995): *What kind of work do you do? data collection and processing strategies when measuring 'occupation' for statistical surveys and administrative records*. STAT Working papers No. 95-1. Bureau of Statistics, International Labour Office, Geneva.

Hussmanns, R.; Mehran, F. and Verma, V. (1992): *Surveys on Economically Active Population, Employment, Unemployment and Underemployment: An ILO Manual on Concepts and Methods*. International Labour Office, Geneva.

International Labour Office (1990): *International Standard Classification of Occupations - ISC0-88*. ILO, Geneva.

Jabine, T. and Tepping, B. (1973): "Controlling the Quality of Industry and Occupation Data". *Bulletin of the International Statistical Institute, 45(3)*, pp. 360-389.

Lyberg, L. (1982): "Coding of Occupation and Industry: Some Experiences from Statistics Sweden". *Bulletin of Labour Statistics, 1982-3*. pp. ix-xxi.

Lyberg, L. and D. Kasprzyk (1997): "Some Aspects of Post-Survey Processing." Chapter 15 in Lyberg et al (1997).

Lyberg, L. et al, ed.s (1997): *Survey Measurement and Process Quality*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc. New York.

United Nations (1989): *International Standard Industrial Classification of All Economic Activities.* Statistical Papers, series M, no. 4, rev.3. United Nations, New York. Geneva.

United Nations (1992): *Handbook of Population and Housing Censuses: Part I: Planning, Organization and Administration of Population and Housing Censuses*. Studies in Methods. Series F No. 54 (Part I). United Nations, New York.

United Nations (1997): *Handbook of Population and Housing Censuses: Part V: Economic Characteristics*. Studies in Methods. Series F No. 54 (Part V). United Nations, New York.

United Nations (1997): *Statistical Data Editing. Vol. 2: Methods and Techniques*. Statistical Standards and Studies - No. 48. United Nations, New York and Geneva.

United Nations (1998): *Principles and recommendations for population and housing censuses*. Statistical Papers; Series M, No. 67. Revised edition.

United Nations,
New York

United Nations & Eurostat (1998): *Recommendations for the 2000 Censuses of Population and Housing in the ECE Region*. Statistical Standards and Studies - No. 49. United Nations, New York and Geneva.

United Nations (forthcoming): *Handbook of Population and Housing Censuses: Part VI: Editing*. Studies in Methods. Series F No. 54 (Part VI). United Nations, New York.