STAT Working Papers

STAT Documents de travail

STAT Documentos de trabajo

No. 95 - 1

WHAT KIND OF WORK DO YOU DO?

Data collection and processing strategies

when measuring "occupation" for

statistical surveys and administrative records

prepared by

Eivind Hoffmann from contributions by

Peter Elias, Brian Embury and Roger Thomas

Geneva, February 1995

INTERNATIONAL LABOUR OFFICE, Bureau of Statistics

BUREAU INTERNATIONAL DU TRAVAIL, Bureau de statistique

OFICINA INERNACIONAL DEL TRABAJO, Oficina de Estadístca

Preface

This *Working Paper* presents practical guidelines for those who want to register the characteristic (variable) "occupation" in a statistical survey or on administrative records. The principles and procedures described are mostly based on experiences from Australia and the United Kingdom, supplemented by discussions which the contributors have had with experts from other countries. The text was originally intended to be part of an ILO manual on the development and use of a national standard classification of occupations. It has, however, proved difficult to finalise the manual, and it has therefore been decided to issue parts of it in this form. A companion *Working Paper* on how to develop the structure and content of a national standard classification of occupations has also been prepared. The two *Working Papers* can be regarded as partial drafts for the manual, and the ILO Bureau of Statistics would welcome all comments on the texts and all suggestions for their improvement.

References are made in some sections to the use of computers when registering information in surveys and for administrative records, as well as when coding "occupation" on the basis of the information obtained. However, most of the discussion refers to a situation where paper and pencil is being used for the recording of the necessary information. This is because most of the principles are unaffected by the use or not of computers, even if the mechanics of the procedures are modified. A further consideration has been the expectation that "paper and pencil" procedures will continue to be the dominant ones in many countries in the years to come.

This *Working Paper* can be seen as a supplement to the **International Standard Classification of Occupations** (ISCO-88) because of the importance of good quality recording and processing of occupational information for the comparability of the resulting data. However, the validity of the guidelines is largely independent of whether or not ISCO-88 has been used as a model for the development or revision of the classification of occupations used for the registration of the "occupation" variable.

This *Working Paper* is based on a draft prepared Mr. **Roger Thomas**, then at the Office of Population Censuses and Surveys, and Mr. **Peter Elias**, Institute of Employment Research, University of Warwick, both United Kingdom. Important contributions have also been received from Mr. **Brian Embury** who over many years has shared with the ILO the results of the path-breaking methodological work of the Australian Bureau of Statistics in the development and statistical use of an occupational classification. On the basis of their contributions Mr. **Eivind Hoffmann**, of the ILO Bureau of Statistics, prepared the text presented here and is responsible for the views expressed and any errors and ommissions.

Comments and suggestions for improvements to this *Working Paper* will be highly appreciated. Please send them to: **Bureau of Statistics; International Labour Office; CH-1211 GENEVA 22; Switzerland. (Fax no: + 4122 799 6957)**.

Farhad Mehran
Director
Bureau of Statistics

Contents

1 Introduction to basic concepts and principles of occupational classifications

The purpose of this paper is to provide a starting point for planning or improving the registration of occupational information in administrative records as well as in statistical censuses and surveys, and the processing of such information. Following an introduction to the basic concepts and principles of occupational classifications in general and the International Standard Classification of Occupations in particular, the topics covered include: the main data collection instruments; methods of evaluating and ensuring the quality of occupational information; strategic and organizational issues associated with coding and processing; the development of coding indexes; the use of computer technology in data collection and processing; and the problems posed by different languages. Its final chapter presents a strategy for mapping from a national occupational classification (NOC) to the International Standard Classification of Occupation (ISCO-88). The paper is summarized through twelve 'golden' rules for the data collection and processing for the measurement of "occupation" for statistical surveys and administrative records.

1.1 What is an occupational classification?

An occupational classification is a tool for presenting information about the types of work which are performed in the jobs found in an establishment, an industry or a country and for organizing this information systematically. It normally consists of two mutually supportive components:

- a descriptive component, which may be just a set of titles of occupations and occupational groups, but which often consists of descriptions of tasks and duties as well as other aspects of the jobs in defined groups. This descriptive component constitutes the dictionary of occupations;

- a classification system, which provides guidelines for classifying jobs into the most detailed groups of occupations and for aggregating these into broader occupational groups.

Occupational classifications can be compared to a system of maps for a country. The top level of aggregation corresponds to a small scale map showing the main rivers, mountains, cities and roads, while the next level corresponds to a set of larger scale maps for each of the main regions, showing smaller towns and the roads between them. At the most detailed level one finds the technical maps used by municipal engineers to plan pavements, traffic lights, road extensions, etc.. These very detailed maps correspond to the detailed job descriptions which are used as management tools by enterprises to organize their activities, formulate wage and salary scales and evaluate jobs. In most countries these are outside the concern of national authorities, except in relation to the management of their own activities.

1.2 The use of national occupational classifications

National occupational classifications and dictionaries are usually designed to serve several operational and planning purposes. Although the detailed occupational descriptions and the classification structure must be seen as an integrated whole, different users have different interests in the various elements. The detailed occupational groups and descriptions are used mainly by client-oriented users, i.e. those responsible for job placement, vocational training and guidance, migration control, and so on - activities with information requirements

similar to those of personnel management. Although they should be designed primarily to meet the needs of such users, descriptions of detailed occupational groups should also include elements necessary for applying relevant classification schemes. The classification structure as a whole is used principally to facilitate statistical description and analysis of the labour market and the social and economic structure of the country, as a basis for public debate and the formulation and monitoring of policies, as well as national and regional development.

The breadth of applications of occupational classifications can be illustrated by the f following examples: Legislators and public sector administrators use occupational statistics in formulating of government policies and in monitoring progress with respect to their application and their results, as well as in manpower planning and the planning of educational and vocational training. Managers and workers' representatives need occupational statistics for shaping working conditions and manpower policies, including compensation, at the establishment and industry level. Psychologists study the relationship between occupations and the personality and interests of workers. Epidemiologists use occupation in the study of work-related differences in morbidity and mortality. Sociologists rely on occupation as an important variable in the study of social differences in life styles, behaviour and mobility. Economists use occupation in the analysis of differences in the distribution of earnings and incomes over time and between groups, and in the analysis of employment and unemployment.

Depending on the purpose of the description or analysis, “occupation" may be the main variable in a statistical study or it may serve as a background or explanatory variable. Used as a background variable, it may serve as a proxy for other variables more difficult or impossible to observe, such as socio-economic group or exposure to certain harmful substances, or it may be used as one element in the construction of other variables, such as social class or socio-economic status.

While a comparison of the distribution of the employed population (or some other variables such as wages, hours of work, work accidents, income, consumption or reading habits) over occupational groups requires that the occupational classification should cover all jobs, the focus in other types of use, whether statistical or client-oriented, is normally focused on specific occupations or groups of occupations. In total the users interested in specific groups may between them also cover all occupations. However, in practice the coverage will be very uneven and likely to exclude jobs in certain areas, especially jobs outside the modern, formal sector.

1.3 The use of ISCO

The International Standard Classification of Occupations (ISCO) is designed to facilitate international communication on occupational information, narrowly or broadly defined, for both client-oriented and statistical users. It should also serve as a basis for developing or revising national occupational classifications, or act as a substitute for a national classification if a national one has not been developed. ISCO must therefore reflect different uses at the national level, while taking into account the special considerations which follow from its international nature.

The main client-oriented applications of an international standard classification of occupations relate to the international recruitment of workers and the administration of short- and long term migration of workers between countries. An internationally developed and agreed set of descriptions for detailed occupational categories which can serve as a common

"language" for the countries and parties involved in such programs, enhances the effectiveness of communication necessary for their execution.

Internationally comparable statistics on occupational groups are used mainly to:

(a) compare the occupational distributions or the distribution of some other variables such as wages, consumption or literacy, in two or more countries;

(b) compare data on broadly or narrowly defined occupational groups in two or more countries, for example the average wages of computer programmers or the number of industrial designers;

(c) merge comparable data from different countries in order, for example to obtain enough observations to analyze the incidence of particular work-related accidents or diseases among workers believed to have similar exposure to harmful substances, or to work under certain relevant conditions.

Experience shows that most users of international statistics need data at high levels of aggregation - usually for type (a) uses. Important exceptions are studies of earnings, work hazards and injuries and other conditions of work which often require detailed data, sometimes cross-classified with industry and/or status in employment.

1.4 Basic concepts and principles of an occupational classification

The primary objects classified in an occupational classification are jobs. A job is defined as a set of work tasks and duties performed by one person or designed to be performed by one person (in the case of unfilled jobs). Jobs which have the same set of main tasks and duties are aggregated (grouped together) into occupations. Occupations are grouped together into narrowly or broadly defined occupational groups on the basis of similarity in the type of work done, i.e. similarity in the tasks and duties performed. The units described in a dictionary of occupations are occupations and occupational groups. (At the enterprise level the wage and salary scheme may describe individual jobs.)

The decisive factors for how well an occupational classification can suit the needs of any user are that the appropriate number of groups are specified and that appropriate criteria are used to define similarity in type of work done. Unfortunately different users have different requirements with respect not only to appropriate level of aggregation but also to the most appropriate similarity criteria. For some users (e.g. insurance companies) important criteria may be whether the work is outdoors or indoors, or whether it requires traveling or not. For others, the social status of the work may be most important, or they may want to focus on the materials worked with, the goods and services produced or whether the work requires direct contact with clients and customers. By deciding on the main similarity criteria to be used in the occupational classification its developers implicitly or explicitly give priority to some users needs over others. The implications of this for the overall use of the classification must therefore be carefully evaluated, but the shortcomings as far as non-priority users are concerned may be significantly reduced if sufficient details can be developed for those parts of the classification which are of importance to such users.

In many of the national occupational classifications that have been revised or developed since 1985, as well as in the latest version of ISCO, ISCO-88, occupations are

identified, defined and grouped mainly on the basis of the similarity of skills required to fulfill the jobs tasks and duties. In ISCO-88, as well as in the NSCOs using the same similarity criteria, two dimensions of the skill concept are used to define groups: skill level, which is a function of the range and complexity of the tasks involved, where the complexity of tasks has priority over the range; and skill specialization, which reflects the type of knowledge applied, tools and equipment used, materials worked on, or with, or the nature of the goods and services produced. With this approach the focus is on the skills required to carry out the tasks and duties of an occupation and not on whether a worker having a particular occupation is, or needs to be, more or less skilled than another worker in the same occupation.

If "Skill level" is used as the main variable for defining the broad distinctions which separate the major groups in the classification, it will by necessity run through the whole classification system. In some national classification systems the “skill level" categories have been defined with reference to the national education and training systems, specifying the type and level of training and experience which new entrants into the occupations are typically expected to have. In ISCO-88 skill level categories are defined by references to UNESCO's International Standard Classification of Education (ISCED); (COM/ST/ISCED; Paris 1976). This does not mean that ISC0-88 assumes that skills can be obtained only by formal education or training. Most skills may be, and often are, acquired through experience and informal training, although formal training may play a larger role in some countries than in others and may also be relatively more important at the higher skill levels. In the ISC0-88 classification system, the decisive factor for determining how an occupation should be classified is the nature of the skills required for the job - not the way in which these skills have been acquired.

'Skill specialization' is used to make both broad and fine distinctions within the major groups of the classification. Specializations are related to subject areas, production processes, equipment used, materials worked with, products and services produced, and so on; and words describing Specializations are used in the titles and descriptions of the more detailed occupational groups. The same types of word are also used to describe the groups of an industrial classification of production activities. For some jobs it will therefore be possible to predict the industry in which they are located with a fairly high degree of success, knowing how they are classified by occupation. It is important to understand that in ISCO-88 and similar NSCOs this is because skills are linked to products, materials, and so on and not because industry is used as a classification criterion.

As job is the primary object classified by an occupational classification, it follows that a person can be classified according to an occupation or occupational group only through his or her relationship with a job. This can be a job held in the past, a current job or a job he or she is looking for. In this context “to have a job” is meant in a broad sense, so that the classification should be applicable to all employment situations: employees, the self-employed and participating family members, i.e. everyone working for pay, profit or family gain. It follows from this that, depending on the circumstances, one person may be classified according to several different occupations if he or she has had (or is expected to have) more than one job. Users who need to work with only one occupation for each person, will have to formulate priority rules for selecting the 'most important' job. In principle such rules are outside the scope of an occupational classification system, but possible criteria for choosing the 'main' job are briefly mentioned in section 3.3.2 “Determining which job to describe".

2 The main data collection instruments

This chapter examines the different sources of information about the occupation of jobs and persons, and assesses briefly the advantages and disadvantages associated with their use, in particular for the production of occupational statistics.

In general job-holders are likely to have fuller and more accurate knowledge about their own jobs than others do. The quality of occupational information is therefore likely to be best if obtained directly from the job holder. However, censuses and surveys, as well as many administrative registrations, may in practice have to obtain the information from someone else. Usually it is another member of the job-holder's household or a representative of the employer who will act as proxy respondent.

The ideal data collection situation is one where a person having both the skills of an interviewer and a coder is in direct communication with the job-holder and able to continue to question him/her until the correct allocation of the job within the classification has been established. In practice this ideal situation can normally only be created when job-seekers are being interviewed for job-placement or work-permits, or in surveys where the interviewers will use a computer assisted coding system to determine the correct code during the interview (cf. section 5.3).

2.1 The national census of population

Censuses of population are very large-scale, multi-purpose operations in which certain general aims, such as operational robustness, timeliness and economy and the achievement of complete enumeration tend to take priority. They give scope for adequate definition of relevant populations of persons, but there may be difficulty in practice in getting enumerators and informants to respond in a way which gives the type of information required by the definitions of 'occupation' and the different occupational groups.

Population censuses may use, or partly use, pre-coded response alternatives for self-completion by the respondents, or use enumerators to question informants and record the information obtained. With both procedures the information written on the forms has to be interpreted by coders. In the first case one or more standard questions and a few instructions and examples are given to assist the respondents directly. In the latter case the enumerators may be given more elaborate, but still limited, instructions related to the questions to ask and the types of answers to seek and record. In both cases the informant is frequently a household member who gives information about his/her job and those of other eligible members of the household.

When using the self-completion method there is generally severe competition for question space on census forms and the mode of questioning must be kept very concise. Each question should be immediately intelligible to every informant; but, given the intention to obtain 100 per cent coverage of the population, it follows that many informants may have poor literacy skills and will be unaccustomed to dealing with forms. Questions which are either long, use "technical" terms or are conceptually complex, or which are accompanied by detailed explanatory or illustrative notes, produce unreliable data because informants cannot or do not make the effort to grasp the implications for the answers they are asked to give. Although the census questions are standardized the degree of control which can be exerted over the quantity and quality of information obtained and recorded is still very limited. This is

particularly so in the case of 'occupation', where on-the-spot quality control is hampered by the fact that there is no simple, rapid and reliable way of recognizing whether or not the information on occupation supplied by an informant is adequate for coding Purposes without actually trying to do the coding.

Checks made in the context of census tests suggest that, even when information is to be recorded on a form left with the household, form fillers seldom consult those whose occupational details they are filling in, but rely on their own knowledge and impressions. In a number of cases this leads to occupational codes being assigned which are different from what would have been assigned if the job-holder had been the informant. Discrepancies occur for two main reasons. In some cases proxy informants are genuinely misinformed about the nature of another household member's job and mis-report it for that reason. Another source of error is that, while having the same job in mind, the proxy respondent may describe it in different terms from the job-holder and thereby cause coders to assign it to a different occupational category.

When the census is carried out by the use of enumerators who fill in a census schedule on the basis of the responses obtained, then the instructions and training received by the enumerators and the time available for recording occupational responses will determine the quality of the information to be used later by the coders. The size of the census operation and the large number of enumerators required normally severely limits the amount of training they can receive on this aspect of the census and the time they can spend in getting adequate information.

The amount of space provided on the census form for the recording of the occupational information, will also be an important determinant for the quality of the information provided, whether self-completion or enumerators are used.

2.2 Data produced by household surveys

In such surveys data will normally be collected through face-to-face or telephone interviews. As with censuses, useful ancillary data for coding occupation are likely to be collected. In principle detailed control can be exerted over the occupational concepts and the form and sequence of questioning to be used. If the information is elicited and recorded in an interview situation using standard questions and supplementary probes (where required), the interviewer should ideally be able to record occupational information in a systematic way and to probe to clarify details. A great strength of continuous surveys is that over time experience can be accumulated and procedures perfected in the fieldwork and coding areas and staff thus brought to a higher and more consistent level of training and competence than normally it is possible to achieve in a census operation. Also, integrated quality control procedures designed to optimize the data collection and processing system as a whole can be developed.

However, cost considerations usually necessitate solutions with less than complete controls. To keep fieldwork costs down and response rates high, large surveys of the labour force may also have to rely on proxy informants to supply data on the occupations of a proportion of the persons covered. This affects data validity to some extent, particularly in cultures where the economic activities of one household member may be largely 'invisible' to others. It may not be possible to train field workers on large scale surveys to recognize where information is inadequate for coding by others, and they may be working under circumstances which rule out lengthy questioning.

2.3 Data from employing organizations

Although employing organizations may maintain personnel records on their staff and their jobs, it is not always straightforward for them to find the information appropriate for the occupational coding of their work force and/or vacant jobs. An additional consideration, particularly relevant for statistical surveys, is that in general, employing organizations may not give the completion of government forms high priority or care, unless it is to their direct commercial advantage to do so. It is thus hard to control the quality, consistency and completeness of data, even if collection is, in principle, backed by penalties for non-completion.

The quality of information provided may be particularly difficult to assess when the employers are asked to provide counts by age and gender as well as occupation on self-completion forms. Then they are effectively requested to classify and jobs and their incumbents. This practice may appear very economical from the point of view of data processing, but it transfers to the form-fillers problems of classification and coding which they are unlikely to handle in a consistent and satisfactory way because they have not received the necessary training and instruction. Problems are also likely to arise where the form in which data are requested does not accord exactly with the form in which organizations keep records. In particular, the occupational classifications which they make use of in their own record-keeping systems may differ significantly from the classification they are being asked to use. The problems associated with this approach can be reduced by tailoring the questionnaire to the industry of the particular employer, i.e. to ask the employer to give the number of employees in each of a described set of occupation groups which are likely to be represented or important in the employer's work force. The set will then vary between firms in different industries. A related approach has been to provide every employer included in the survey with a complete dictionary of occupations which describes the possible occupations and their codes.

It seems likely that better quality data may be obtained by relieving the employer of the coding task altogether, for example by asking him/her to provide anonymous individual information about randomly selected individual members of the work force, including the job title and a short statement about main tasks and duties which can be used by the statistical office as a basis for coding the occupation following the procedures outlined in chapter 3 below. This approach gives the possibility of collecting information about a wider range of characteristics of the selected individuals, such as age, wage rates and earnings, years of employment and qualifications, thus providing possibilities for much richer statistics.

When collecting occupational information from employing organizations, it is important to recognize that, depending on the size of the employing organization the person giving information about a particular job within it, may or may not know personally the tasks of the job. Normally he/she will have to rely on general knowledge about the firm or unit and on the administrative records available. Such records have normally been designed for other purposes and the designations and descriptions used may serve to reflect e.g. location on the salary scale, not differences in tasks. Inadequate or misleading information may therefore easily be given, especially if the organization has worked out a cheap and easy way of making a return - i.e. one which minimizes time and effort while being more or less complete without obvious internal inconsistencies.

2.4 Data produced by administrative procedures

In general, the quality of occupational information from administrative procedures

depend entirely on the administrative agency's dependence upon their quality and/or its concern with the use of the information to produce good quality statistics. It is therefore important to distinguish between those administrative registrations which produce information about jobs because this information is needed for the work of the agency and its officers, and those registrations where occupational information is collected but play no real role in the decisions to be made. Examples of the former are the information recorded in agencies charged with finding workers for vacancies and employment for job-seekers, and when work-accident insurance rates depend upon the type of tasks (and risks) of the job. Examples of the latter may be the recording of occupation on tax forms or on forms registering births, marriages and deaths.

Clearly, even when used by the administrative agency only those distinctions which are relevant to the decisions are likely to be recorded with any precision. The quality of the recorded information will also depend on whether mistakes in the initial recordings are in fact corrected on the files when they are discovered. In many cases such mistakes may not have an influence on the decision, and even if it has, only the decision may be changed and not the incorrect information. One example is the case of job placements: If a satisfactory job (candidate) is found for a job seeker (vacancy), then an original mistake in the occupational code is not likely to be corrected, as to do so will have no "practical" usefulness.

There may also in some cases be an incentive for the informants to give misleading information to the administrative agency, for example if the granting of immigration or emigration permits depends on skills or occupation. An illustrative historical example concerns the convicts transported from Britain to Australia in the early 1 800s: Because the government had priority on the use of skilled craftsmen, but private employers gave better pay, convicts with craft skills had a strong incentive to downgrade their past occupational experience when asked about it before landing in Australia.

3 Evaluating and improving data quality

This chapter describes the concepts of validity and reliability as applied to the occupation variable and draws attention to certain key conventions and assumptions which underlie the classification process. Methods of assessing validity and reliability, together with techniques for ensuring the quality of information about the occupation are discussed. Particular attention is given to formulation of the questions used for data collection and to the coding process.

3.1 Defining data quality

3.1.1 Validity and reliability

A *valid* occupational code should convey accurately into which category of the classification the pattern of tasks present in a job best fits, and should reflect adequately the distinctions between differences in type of work. An occupational code assigned to a job is *reliable* if repeated application of the same data collection and data processing procedures to the same case leads to the job being assigned to the same occupational category, independently of the particular job incumbent or respondent and the coder.

A data collection and processing system which delivers high validity must also have high reliability - that is, it will consistently place jobs in the correct occupational category. Low reliability sets limits to validity - that is, if the system does not work in a consistent way it cannot classify jobs in a way which is consistently correct. However, high reliability need not imply high validity of the results. For example, a superficial coding procedure might achieve consistency, but only through ignoring important parts of the data recorded or inappropriate use of certain "catch-all" categories. This is an important point to recognize, because whereas validity is the more fundamental criterion of data quality, reliability/consistency is easier to measure and optimize, and may therefore sometimes be aimed at. at the expense of validity.

3.1.2 Assumptions and conventions

The practical work to classify jobs and persons by occupation depends on certain key assumptions and conventions. These are necessary for practical reasons, but in certain cases the assumptions may be invalid or the conventions may oversimplify or distort reality and thereby lead to results which are less valid and/or reliable than wanted.

a. Conventional definition of a 'job': In practice the work to classify jobs by occupation tends to be based on the assumption that there is a single, integrated and stable pattern of activities and tasks which characterizes the work which that individual does, and that the basic information from which he/she should be allocated to an occupational group reflects this pattern. This may not always correspond to the normal understanding of 'a job', namely as an area of work which a person is separately paid for. Censuses and surveys, when faced with an individual who has more than one relevant job, often simplify by focusing occupational questioning on 'tine main job', usually defined as the one to which he/she devotes most time or through which he/she earns most money. Registrations in employment offices may tend to focus on the last job held and/or the type of job the client is looking for. A balance has to be struck between the consequent loss of information and the additional complexities and sources of error which result from attempting to cover multiple jobs. Focusing upon 'the job' as the unit of reporting for occupations may also

create distortions where one job involves the exercise of several different sets of occupational tasks (e.g. both managerial and professional). In general, practical occupational classification procedures have been to give priority to one or other set of occupational tasks. One implication is that statistics derived through classifications can be used only very carefully, and selectively to estimate how many persons are currently exercising particular occupational skills.

b. Assumption of consistent usage of job titles (and supplementary information) to describe job characteristics: The degree of discrimination and consistency with which job titles are normally used sets limits in practice to the reliability and validity of the occupational codes in conveying information about the actual distribution of job activities and job skills in the population. Occupational coding practice depends to a large extent upon recognition of job titles. Even if supplementary information is specified as being required, the starting point is normally a job title. It is implicitly assumed that each job title (or set of synonymous and specialization job titles) is used by informants in a consistent way to denote a particular pattern of job activities and skills, and that this differs from the pattern of activities and skills denoted by each other job title (or set of titles). However, in practice this assumption is only approximately satisfied - i.e. job titles in many areas of employment are not used very precisely and consistently. Thus the patterns of activities and tasks denoted by several different job titles may be almost the same, while the same job title may denote a spread of patterns of activities and tasks which at the extremes are quite different from one another. This is because job title usage is in fact affected by a variety of factors other than the current task and activity content of the jobs concerned, e.g. because of the conventions adopted by particular payment systems, or because usage preserve historical distinctions which have become invalid through technological or organizational change while in new or changing areas of work terms which are used in a consistent way may not yet have emerged. In spite of these problems the job title is by far the most valuable single piece of information practically obtainable for purposes of occupational classification. Efforts to detect and, if possible, correct for gross violations of the assumption of consistent job title usage must therefore be made part of the coding procedure, as discussed in chapter 4 below.

c. Convention of unique classification of each job: It is a universal practical convention that each 'job' should be allocated to one and only one basic category of the occupational classification structure. However, in practice it is often clear that a job as described may fit into more than one category, either because the information is too brief and vague to be used to identify a unique occupational group or because the particular combination of tasks of the job is not precisely reflected in the classification. In the former case it is likely that more, and relevant, information would be useful, but this will not be the case in the latter. Both cases require the formulation of clear rules to guide the coders, cf. section 2.4 below.

3.2 Monitoring and evaluating data quality

3.2.1 The monitoring and control of data quality

Error in the classification of occupational information is a universal problem which is difficult to limit to acceptable levels. All statistical agencies which produce occupational statistics therefore need to monitor and control both validity and reliability of the output. The simplest way to monitor data quality is to look for net bias in the statistical distributions produced. However, really adequate quality control requires close monitoring of the collection

and coding process to detect gross classificatory errors:

a. Net bias in distributions: The presence of errors may become obvious when implausible occupational distributions occur or when impossible or unexpected differences between surveys or fluctuations over time are detected. In such cases the source of error generally cannot be diagnosed with any certainty without close examination of the data collection and coding procedures which have produced the results. The results of such investigations often prove to be that the extent of net bias is small compared with the amount of gross classificatory error which underlies it. This is because many errors in the allocation of cases to categories (gross classificatory error) cancel each other out when univariate distributions are considered.

b. Gross classificatory error: The best way to detect the existence of large amounts of gross classificatory errors is to collect and code independently data relating to the same set of cases and compare the outcome. It can and does happen that two distributions for the same set of cases, which are not very discrepant in net terms, produce a case wise coding consistency rate of only 60 per cent or lower. The presence of such levels of gross error can have a serious distorting effect where occupation is cross-classified with other variables such as sex, age or income.

3.2.2 Assessing data quality

In order to assess the validity of occupational data in a way which lends itself to improving the quality of the data, it is necessary to review the stages of data transmission and transformation which occur between the initial approach to an informant and final allocation of an occupational code. Errors and distortions may occur at each stage. For example:

a. the informant may mis-report job title/job activities, e.g. because of genuinely misunderstanding the question, to inflate apparent status or skill level of job, to conceal illegal or socially unacceptable activities; etc.;

b. the informant may give a vague or ambiguous answer or use occupational terms in a sense or in an order different from that assumed in the coding index or instructions;

c. a proxy informant may misrepresent through ignorance the subject's job title or activities;

d. questioning may fail to elicit key aspects of job for coding purposes;

e. a field worker may misunderstand, misrecord or omit essential details of what informant says;

f. a coder may misread or misunderstand what is recorded;

g. a coder may make errors in applying procedures to arrive at the occupational code.

Except in special cases where good alternative sources of information exist, there is only one satisfactory and comprehensive way of monitoring the correctness of occupational data - i.e. of detecting, identifying and assessing errors and distortions arising from any of the above causes. This is to repeat the process of collecting and processing the data, using the best and most thorough methodology available, so as to establish a validation criterion. This should be done in a way which as far as possible enables errors arising from the different sources listed above to be separately identified. The resulting information on the absolute and relative importance of the different sources of error can then be used to guide quality control measures- for example the allocation of resources between efforts to improve data collection

and efforts to improve coding.

Quality control exercises need to be done on a sufficiently large and statistically valid sub-sample of the census, survey or administrative data collection process to enable useful conclusions about the correctness of the original data set and procedures to be drawn. If separate results on the validity of small occupational categories are required, the control sub-sample will, of course, need to be very large or especially designed with this in mind. However, more aggregate measures are also useful. Two studies done in the context of the 1981 UK Census of Population provide examples of this:

One study was a follow-up of a sample of households for which the census form filler (any household member) had provided occupational details. An interviewer asked each adult member of the household to give his/her job title and description; these data were then re-coded and the result compared with the code which had independently been assigned to the information recorded on the census form. The detailed occupational code (from a frame containing about 350 categories) assigned on the census agreed with that assigned in the follow-up in about 74 per cent of cases. When codes were aggregated into six broad occupational groups the rate of agreement rose to 87 per cent. Of the 26 per cent of cases where there was disagreement at the detailed level, about 5 per cent arose because the coders assigned different codes even though the recorded information was virtually the same (that is, through unreliability of the coding process - see sources of error f. and 9. above). However 21 per cent arose because fuller information was recorded at the interview than on the census form (i.e. error arose from some combination of sources a.-e. above). Agreement levels were particularly low for residual categories such as 'Other general labourers'.

The other study involved a comparison between the occupational codes assigned to data collected (for the same persons) by the Census of Population and by the Labour Force Survey, which is conducted by interview with a household informant. Here the agreement rate at a detailed level of coding was 70 per cent.

The level of consistency at which occupations were classified in the UK census may be lower or higher than the rates achieved in censuses elsewhere in the world, but the presumption is that this is fairly typical of censuses carried out using household form fillers and central coding. The results indicate that possibly a fifth of the codes assigned may be incorrect where the criterion of correctness is exact allocation to categories of a detailed occupational coding frame. However, there is evidence that this performance can be improved upon by using better trained coders and stricter and more comprehensive coding rules and supervision than has been normal in the processing of censuses and large surveys. The increased costs of such improvements in coding quality can probably only be offset by the use of Computer Assisted Coding (CAC), cf. section 5.3. below.

Problems of unreliability affecting occupational data can arise at broadly two stages, namely, during data collection and during data processing: At the data collection stage the issue may be posed as: what is the probability that field workers instructed to apply the same procedures for collecting occupational data to the same individual or household informant in the same objective circumstances would elicit and record verbatim responses which were exactly equivalent for coding purposes? Studies of the reliability of field procedures for collecting occupational data are rather rare. When the costs of re-contacting informants are incurred the purpose is usually to set up a validity criterion by using improved data collection procedures, rather than to go through exactly the same procedures again, in order to assess their consistency. There are also technical problems such as memory contamination of responses. The evidence which does exist suggests that job-holders normally refer to their

jobs in the same terms when asked on separate occasions, using the same question wording, but that proxy informants may use different terms. The amount of information given in answer to questions about job activities varies more according to the interview context and procedures. In particular, interviewers make varying judgments about whether or not it is necessary to probe for additional information and in the exact probes which they use. As a result of all this, different information will be available to coders, depending on which interviewer and which informant was involved. From the point of view of the informants the information given in response to two approaches may appear equivalent or complementary rather than contradictory, but, because of the nature of the coding process, the result often is that different codes are being assigned. The proxy informant effect is probably quite substantial. A comparison carried out in the UK of the data produced by questioning spouses separately about each other's occupations showed that the detailed occupational codes assigned differed in 14 per cent of cases.

At the data coding stage the reliability issue may be posed as: would coders faced with the same written raw material and instructed to apply the same classificatory structure and coding instructions arrive at the same code? The reliability of the coding process is easier to evaluate than that of the data collection process because it lends itself to controlled studies in an office environment. Tests of coder reliability are best done on a 'blind' basis - that is, with each of several coders assigning codes to the same set of raw material without knowledge of the codes assigned by the others. Results of checks of this kind carried out in the UK and elsewhere using raw material from surveys and censuses suggests that average agreement rates amongst trained coders using a classification containing several hundred occupational categories are within the range of 85-95 per cent. Experienced coders tended to perform somewhat more consistently than those with little experience. As might be expected, agreement rates are depressed where many of the responses to be coded are vague or lacking relevant details, in particular with respect to the use of certain residual or 'dustbin' categories.

The performance can be improved upon by using better trained coders and stricter and more comprehensive coding rules and supervision than has been normal in the processing of censuses and large surveys. The increased costs of such improvements in coding quality can probably only be offset by the use of Computer Assisted Coding (CAC).

3.3 Question design

While it is probably impossible to make the behaviour of coders entirely consistent and error-free, even when they are provided with 'good' raw information, much the largest source of error lies in shortcomings of the verbatim raw material as elicited and recorded in the field. In order to obtain valid occupational information which can be readily and reliably coded, it is necessary to ask for several separate items of information. Each item supplied may contain errors or shortfalls, these being broadly of the following types:

a. Information given is specific but incorrect because of informant's ignorance of facts.

b. Information given is specific but incorrect, because of informant's misunderstanding of what is required.

c. Information given contains insufficient detail about job for accurate coding.

d. Information given is misleading or ambiguous.

Defect a. can in general only be remedied by using better-informed persons as

informants. Avoidance of defects b.-d. depends essentially on good design of data collection documents and procedures. However, the means of achieving this are not at all easy or obvious.

3.3.1 General problems of question formulation

It should be kept in mind that most of the systematic experience for the following advice on question formulation (and coding), have been collected in industrialized countries using English as the first language. More limited and partly less systematic information about the experience of other countries does not, however, indicate that the conclusions drawn from this experience must be drastically modified to be valid for the circumstances of other countries, even though the ways in which terms and concepts are used to refer to work, jobs and occupations vary between cultures and languages. This reflects differences in economic and employment structure; differences in what is commonly seen as ‘work' and 'economic activity'; differing linguistic idioms; and so on. Therefore, *uncritical* translation of set questioning formulas and labour market or occupational concepts from one country or language to another is a recipe for disaster. The designer of forms and questions needs to be familiar with the particular cultural/industrial/employment universe with which he/she is dealing; and also with the current usage of occupational terms and concepts amongst persons who will be required to answer the questions. It is only with such familiarity, *based on practical experience,* that it is possible to translate the required concepts into words and questions which informants are likely to recognize and to interpret consistently in the intended ways.

The goal to be aimed at is: simple questions using familiar, widely-understood terms and concepts, which do not require special explanation. The longer and more complex the instructions provided on forms or to the interviewer/enumerator, the less likely they are to be read, understood and acted upon, particularly by members of the public acting as form fillers. Form fillers, also trained interviewers, are in any case reluctant to follow instructions which run counter to their natural way of behaving, thinking and expressing themselves. Unfortunately there is always a gap between the vagueness and flexibility with which many terms are used in common speech and the exactness and rigour which is ideally required by administrative or statistical use. The question formulator must constantly judge how, given the circumstances for collecting the information, to best obtain from the respondent the information required to code with the required standards of exactness. Complete avoidance of error and inconsistency is not a realistic goal; an approximate but simple and easily understood approach may perform better than a technically exact but wordy and complicated formulation which may be difficult to "understand and implement.

3.3.2 Determining which job to describe

Before asking for information about a person's job or occupation decisions must be made with respect to whom to ask and about which job(s) to ask. If in a statistical survey we want to know about current jobs, then we can only ask persons who currently have a job. If we want to know about past jobs, then we may ask everyone who had a job in the relevant period. If we want to know about wishes for possible future jobs, then we may want to ask people who are currently looking for work. We may also want to know about the type of job which persons have been trained for or have experience from, if the purpose is to register the type of qualifications for persons seeking work.

The following are three examples of 'jobs' for which occupational information is often asked in censuses and surveys:

a. 'the person's main job last week';

b. 'the most recent full-time job';

c. 'the person's usual occupation'.

Example a. is the job concept most commonly used for those persons who are defined as being 'employed' according to the international definitions of 'current employment'. It is also close to the concept of 'my job' held by most persons in regular full-time employment in the formal sector in industrialized countries. In focusing exclusively upon a 'main job', question formulators implicitly assume that, for the great majority of employed persons, a main and regular job can be identified in a straightforward way. In some countries and for some groups such assumptions are not justified. To ensure that consistent answers relating to 'main job' are obtained from persons who had more than one job last week, a definition is needed. 'Main job' may be defined as, for example, the job providing the highest proportion of earnings, or as the job to which the person devoted the largest amount of time.

The 'main job last week' is not applicable for persons who were not in a paid job last week - for example those defined as unemployed. Both the type of job they were looking for and their job experience are of interest to placement officers and potential employers as well as to users of the statistics. In practice most censuses and surveys will only ask for occupational information relating to 'most recent main job', i.e. concept b.

Some users of occupational statistics find it inappropriate for their needs to use concepts a. and b. as basis for determining a person's occupation, because they do not provide a complete picture of the person's occupational skills and or experience. These users typically would like to use 'occupation' as an explanatory variable in the analysis of differences in lifestyle, i.e. consumption and time-use, and work-experience as well as social mobility and differential morbidity and mortality. For such studies the most appropriate job concept would be c., taken to mean the type of job for which the person was trained and/or in which he/she had most of his/her work experience. To obtain occupational information of this kind will in a large number of cases require both more information about the person's occupational career than it is normally possible to collect in population censuses or labour force surveys, and judgment as to which of several occupations is the person's 'usual' occupation. For these reasons occupational information about job concepts such as c. are normally only collected in specially designed surveys. Such information about job seekers may be collected by employment services to provide a basis for the search for appropriate vacancies.

3.3.3 Questions and related instructions

The purpose of the questions about his/her occupation is to stimulate the respondent to give to the interviewer, and therefore also to the coder, the type of information needed to determine the best occupational code to be given to the job(s) described by the respondent. The purpose of the related instructions is to enable the interviewer to probe for more adequate information if the initial response does not provide the basis for unambiguous coding. The instructions may be specific, i.e. request the interviewer to always use specified follow-up questions to particular responses, or they may be more general and explanatory and leave the exact formulation of probes to the judgment of the interviewer.

In most censuses and surveys the questions designed to obtain occupational information will only be asked after the responses to other questions have established that the respondent has, has had or will (want to) have a job which can be described. Questions asked

about the job normally cover the type of activity (industry) and the institutional sector of the employer or firm, the type of job held by the respondent and the status of the respondent in the job (as self-employed, employee, member of a producers' cooperative etc.). There is reason to believe that the sequence in which these questions are asked may influence the resulting answers. However, little concrete evidence is available as to the nature and size of the influence. The most common sequence of questions seems to be: industry, institutional sector, occupation and status, and it is therefore recommended to follow this sequence unless there are strong arguments for changing it. The best would be to experiment with different sequences of questions. It is, however, difficult to specify precise criteria for evaluating the results of such tests.

Having established for which jobs one wants to collect occupational information for the different population groups covered by the survey the question formulator has to decide what questions to ask and what instructions to give to the respondents and the interviewers/enumerators. Because the precise questions and instructions to interviewers/enumerators as well as to respondents will depend on how and by whom it is intended that the coding of the response is to be carried out, this has to be clarified before the question formulator can decide on the precise formulations. In practice we have the following three alternatives for how the coding should be carried out:

a. By the respondent himself/herself to a pre-defined group;

b. By the interviewer/enumerator during the interview or before the questionnaire/schedule is forwarded for further processing;

c. By specially trained coders as part of the processing (i.e. data entry and consistency control) of the questionnaire/schedule.

The advantages and disadvantages of these possibilities are discussed in section 4.2.2 below.

Having chosen to let the respondent do the coding the following question/instruction is often used:

What is your (his/her) occupation? (Please mark the most appropriate group.)

One weakness of this formulation is that the term “occupation" may not be well understood by many respondents. Another is that no reference is made to any particular job, and the respondents may therefore tend to give responses which are more related to how they see themselves than to what they are doing in their main job or in their last job - i.e. the response may be for a job concept of type c. (as described in 3.3.2 above), rather than a. or b. A better formulation may therefore be:

Based on the main tasks and duties in your {his/her) {main/last) job, which type of job would you say it is/was)? {PIease mark the most appropriate group.)

The best guidance to the choice which the respondent has to make is to label the alternatives as clearly as possible with respect to their respective coverage. It is important to note that the labels used need not be the official titles which the groups may have been given in the occupational classification. If space is available on the forms/cards used the official titles may be supplemented or replaced with text which better explains to non-specialists the scope of the groups and the distinctions made between them. If it is intended that the respondent's choice should be made in the presence of and with the guidance of an

interviewer, the latter should of course be given more detailed instructions on the proper understanding of the groups, especially as concerns jobs which can be considered to be close to the dividing lines between them. Specific advice on this point can only be given in the context of a concrete set of occupational groups and the national situation.

When someone other than the respondent is the coder of the occupational response, then the questions used must be able to obtain answers which enable the coder to find the most appropriate group for each job. Experience from many countries demonstrates that in many cases the most appropriate response is just one or two words - the job title. However, more commonly a few words to describe the main tasks and duties of the job are needed to supplement the job title. The following examples can be used as illustrations:

United States, Census of Population (1980):

29. OCCUPATION

a. What kind of work was this person doing?

b. What were this person's most important activities or duties?

United Kingdom, Population Census (1981):

1 2. OCCUPATION

a. Please give full and precise details of the person's occupation.

If a person's job is known in the trade or industry by a special name, use that name. Precise terms should be used.

b. Please describe the actual work done.

Australia, Population Census (1991):

OCCUPATION

a. In the main job held *last week* what was the person's occupation?

Give full title

For example: accounts clerk, civil engineering craftsman, fast foods cook, floor tiler, extruding machine operator
For Public servants: state official designation as well as occupation.
For armed services personnel: state rank as well as occupation.

b. What are the main tasks that the person *himself/herself* usually performs in that occupation?

Describe as fully as possible.
For example: recording accounts, preparing drawings for dam construction, cooking hamburgers and chips, fixing cork tiles, operating plastic extruding machine.

The sample questions given above all pre-suppose that the form is being completed by the respondent or a member of the respondent's household. If interviewers are used, then instructions and examples are more appropriately provided during the training of interviewers and in the interviewers' guidelines for the conduct of the data collection.

There is some debate about whether or not to include instructions, prompts, examples etc. as attachments to the questions on forms designed for self-completion. Some argue that the questions should be short and simple to encourage the respondent to read the whole question and provide some type of answer as opposed to none at all. It is also argued that the inclusion of examples will bias the responses obtained. Others argue that instructions and examples can clarify the meaning of the question and help to elicit useful information. It is argued that many respondents do not always have a very clear concept of terms such as occupation and industry and that examples help to explain the type of information which is required for coding purposes. There seems to be some evidence that the former approach works best for persons with a limited amount of education and experience with written material, while the latter approach works best with respondents with more education and a higher degree of literacy.

The instructions and examples included with the questions on the Australian census form were carefully chosen to target specific problems often encountered when coding occupation. They were developed after an analysis of thousands of occupational responses obtained in previous population censuses and census tests. Responses to questions with and without instructions and examples were compared for their utility in coding and possible bias in the replies received. The instructions and examples were refined on the basis of extensive field testing. All example occupational titles consist of at least two words; specific instructions are given to public servants in an attempt to obtain a meaningful job title rather than merely the response 'public servant'; defense personnel are targeted in the same way as the Australian classification classifies occupations in the armed services together with their civilian equivalents whenever such equivalent occupations exist. Similarly, the tasks listed in the examples are used to elicit more detailed information for vague but common occupational titles such as clerk, craftsman, cook, machine operator. The italicized *himself/herself* was added to the task question after experience in the 1986 census showed that many respondents were describing the activities of the establishment rather than the tasks of the worker.

The results of the Australian development work showed that the quality of the responses bore a direct relationship to the amount of room provided on the form for the respondent to write the answer. It appears that respondents take the size of the space provided as a cue to the level of detail required. The 1991 Australian census form provided five lines for the occupational title response and six lines for the task response. The area provided to record the response to each question was 40 mm wide by 35 mm long.

In practice, most census respondents and form-fillers give what they regard as the normal title of the person's job in reply to question (a) in the above examples and a very limited amount of detail on activities in reply to question (b). Instructions to give 'full and precise details' of jobs are in themselves of limited effectiveness in eliciting readily codeable data, both because of the limited space on the schedule or questionnaire for writing responses and because the key information needed to code correctly will not invariably form part of the job title or, indeed, of what the form-filler considers to be an adequate description of tasks. The nature of this key information varies in different parts of the classification. For example, in certain cases the material with which the person works is a key item, but in other cases not; in some cases the mix of managerial versus other activities is important, but not in others. However, the training of form-fillers should provide them with guidance on the type of information which will be useful in various cases. There may be more scope for such training in the case of continuous labour force surveys or other sample surveys conducted by a relatively stable and well-trained field force than in the case of censuses, and also in the case of placement officers in the empIoyment services.

The differences in question formulation between the three examples may not significantly influence the reliability and validity of the responses. Explicit instructions regarding which jobs the questions related to preceded the US and UK questions, while the Australians specified this in the question itself.

Experiments in a number of countries with different question formulations seem to indicate that the intended meaning of the term 'occupation' may not be easily understood by all respondents, that reference should be made to the job of the respondent when the occupational questions follow those on the industry of the workplace and that reference should be made to the usual or main tasks and duties. This means that the following formulation may be a good starting point for experiments to find the most effective questions given the national circumstances:

What type of work do you (he/she) do in your (his/her) job? What are the main tasks and duties n this job?

This formulation assumes that the job one is referring to is defined by previous instructions or questions.

The ideal occupational response consists of both a clear occupational title and a few specific words on main tasks and duties. The following are some examples of good responses, which should be easy to code to most national occupational classifications based on the same principles as the International Standard Classification of Occupations (ISCO-88):

Farm labourer, picking peaches';
Labourer, digging ditches
Labourer, carrying and tidying on site
Street vendor, selling trinkets and jewelry'
Night-watchman, walking grounds and inspecting
Cleaner, washing dishes
Hairdresser, doing ladies' hair
Taxi-driver, driving taxi
Fork lift driver, shifting, loading and unloading goods
Assembler, putting together calculators
Welder, building and repairing ships
Machine-operator, controlling wine bottling plant
Baker, making bread and pastries
Baker, operating bread producing plant
TV-mechanic, repairing TV sets
TV-technician, controlling broadcasting of TV signals
Nursing assistant, making beds, serving food
High school teacher, teaching maths and physics
Driving teacher, giving driving lessons
Night stocking clerk, restocking supermarket shelves
Night filler, restocking grocery shelves
Hospital chaplain, care of terminal patients

The interviewers should be instructed on how to probe for more specific information when they receive vague responses like:

Clerk, clerical work
Civil servant, office work
Public servant, grade 5, administration
Farm worker, farm work

Salesman, selling
Labourer, manual work

Information about the activity of the workplace of the respondent will in many cases be necessary to resolve ambiguities in coding the occupation. Also for this reason it is important to ask the questions concerning the activities of the workplace as effectively as possible. In the three surveys referred to above the industry questions were the following:

United States, Census of Population (1980):

28. INDUSTRY

a. For whom did this person work? (Name of company, business, organization or other employer)

b. What kind of business or industry was this? (Describe the activity at location where employed)

c. Is this mainly (fill one circle):

Manufacturing;
Wholesale trade;
Retail trade; Other (agriculture, construction, service, government, etc.)

United Kingdom, Population Census (1981):

11. NAME AND BUSINESS OF EMPLOYER (IF SELF-EMPLOYED THE NAME AND NATURE OF THE PERSON'S BUSINESS)

a. Please give the name of the person's employer. Give the trading name if one is used and avoid using abbreviations or initials.

For members of the Armed Forces, civil servants and local government officers see notes on back page

b. Please describe clearly what the employer (or the person if self employed) makes or does. For a person employed in private domestic service write 'Domestic service'.

Australia, Population Census (1991):

INDUSTRY

a. For the main job held *last week*, what was the employer's trading name?

For self-employed persons, print name of business
For government employees, print full name of department and division, branch or section.
For teachers, print name of school

b. For the main job held last week, what was the employer's workplace address?

For persons with no fixed place of work, provide address of depot or office.

c. What kind of industry, business or service is carried out by the employer at that address?

Describe as fully as possible, using two words or more. For example dairy farming, footwear manufacturing.

One reason for asking about the name and address of the employer (business) at which the respondent works is that the name often gives an indication of the type of business and that many statistical agencies maintain a business register with the correct industry codes which can be transferred to the respondent if he/she can be matched to the correct business in the register. The 'type of business' question is of course for all the cases where such matching is not possible.

When formulating the industry question it is important to remember that the 'industry' is used to classify businesses and places of work on the basis of their main type of products produced, services provided or function. A job can be given an 'industry' code through its relationship to a workplace or an employer, and a person through his/her relationship to a job. One should try to avoid using the word 'industry' because in many languages it is used in every-day speech to designate one particular industry group, namely 'manufacturing'. Based on these considerations, the following may be a reasonably good starting-point for the testing of alternative formulations of the industry question: "What type of products, services or functions are made or provided by your (his/her) place of work?

Procedures for the coding of 'industry' will not be discussed in detail, but many of the general points concerning the coding of 'occupation' can be applied equally to 'industry'. However, it should be mentioned that there is a special problem concerning the 'industry'-coding of persons working for temporary work agencies or who are seconded from one employer to the workplace of another. In these cases the industry code for the employer-may easily be different from that of the workplace. But, only if the workplace is chosen can 'industry' be validly used as information for the coding of 'occupation'. However, some users of statistics on 'industry' may prefer that the industry of the employer is chosen. The best would of course be if both codes could be recorded.

4 The development of a processing strategy

This chapter discusses the aims of an occupational data coding and processing strategy, the main strategic choices to be made and the main organizational factors determining the dimensions of the coding and processing tasks. Whilst recognizing that requirements will vary from country to country, this section covers many of the common issues which must be addressed.

4.1 Objectives

As outlined above the main aim of the occupational coding and processing of information about the occupation of a job is to determine and record, from the responses obtained from the respondent, to which of the groups in the classification the job of the respondent belongs, while retaining as much as possible of the information contained in the responses. This task has to be completed within an overall processing plan, to a pre-specified timetable and either within pre-specified cost or in a fashion which will minimize cost given the specified data requirements. The development of a processing strategy with these aims needs to consider many aspects and requirements of the processing task:

(a) the existing data processing establishment and infrastructure;

(b) whether the task is continuous, recurrent or ad hoc;

(c) the number, type and format of the information to be processed;

(d) the volume of data to be processed and the throughput rates required;

(e) whether the processing of information about occupation is embedded within a much larger and more complex data processing task;

(f) the arrangements for entering coded data on to a computer medium.

The precise impacts of these aspects will depend on the choices made with respect to some strategic aspect of the coding and processing. These aspects will be discussed in the following section.

4.2 Strategic coding and processing options

4.2.1 Process 100% of cases or a sample only

Occupational coding is typically one of the most expensive and time-consuming operations in processing a census or survey. In order to reduce costs, make the management and quality control of coding easier and enable results to be produced earlier, consideration should be given to obtaining/coding occupational data for a sub-sample only.

This policy may be implemented at the data collection stage by fielding longer and shorter versions of the census or survey form, so that only a sample of the population (or sub-sample of the survey sample) supplies occupational information. An alternative which avoids error-prone field sampling procedures and administration of different forms or schedules is to collect the information from everyone, but to process it for a sample only.

In a population census sampling implies the introduction of sampling error, which will

be an important concern when estimates are to be produced for small population groups or small geographic areas or small groups. As the provision of small area and small group data on a nationally consistent basis is seen as one of the major functions of a population census in many countries, many users may see the processing of occupational information for a sample only as contrary to the role of the census, and may decide that the occupational information may just as well be dropped from the census altogether. This will especially be the case if there exists a regularly conducted, e.g. annual, labour force survey which already does provide occupational information on a sample basis.

Other considerations relevant to the coding of a sample only (or a sub-sample) are the following:

a. Census confidentiality may preclude the production of detailed occupational distributions for small areas. Then relatively little may in practice be lost by sampling since detailed occupational statistics cannot be produced for such areas.

b. Sampling may involve a serious loss of analytic power if summary socio-economic indices for small areas are to be produced using data on occupational distributions.

c. In a sample survey the use of a sub-sample for the coding of 'occupation' and related variables will mean that sampling errors will increase for all groups, making inter-group and over time comparisons more difficult.

Attention must be given to the design of the sample, balancing the requirements of statistical precision against those of operational simplicity and robustness. The sampling fraction should in principle be determined by balancing the precision of estimates required for the smallest aggregates of the population for which separate figures are to be produced, against the saving of cost and time resulting from reducing the processing load. If the sample is to be selected by hand it is also important that the sampling procedure should be simple (e.g. select every n-th case, with n = 5, 10, 15, 20 as appropriate). The sampling should be done on a probability basis. In practice this can conveniently be approximated by taking a systematic sample of census households across each data collection area (enumeration district) and including all members of those households in the processing sample. In this case, due to variable household size and to the clustering of individuals within households, this will result in a departure from a true simple random sample of individuals (the units of analysis for occupational statistics) and one should try to estimate this sample design effect, using appropriate parameters and formulae.

If the processing sub-sample is drawn from completed forms received, but the population base includes a significant proportion of households for which no completed household form (containing occupational details) was received, the sampling interval to be applied to forms may need to be adjusted so as to achieve an exact and convenient 'grossing up' or weighting factor (e.g. 1 in 10).

In theory differential sampling schemes may be considered, which are designed to provide more detail on some sub-groups of the economically active population than others. In practice, however, decisions on sampling must be strongly influenced by the need to avoid making the processing strategy too complicated. When a decision is taken to process certain data on a (sub-)sample basis only, a number of different census/survey questions and variables is likely to be affected. The optimum sample size and design is likely to vary for different variables, but it will be important to keep the number of different data processing options to a minimum. This is likely to dictate the use of one sample for all sample processing purposes, giving just two separate sets of processing procedures (often referred to as 'data

streams').

4.2.2 Field or office coding of occupation

In section 3.3.3, reference is made to the choice of coders, and it is mentioned that with surveys and censuses there is a choice between:

a. the respondent coding himself/herself to a pre-defined group;

b. the interviewer/enumerator coding during the interview or before the questionnaire is forwarded for further processing;

c. specially trained coders as part of the processing (i.e. data entry and consistency control) of the questionnaire.

These alternatives are also present when planning the procedures for capturing and processing information about "occupation" in administrative registrations. The choice is a strategic one with respect to balance between costs, quality and control of the coding process.

4.2.2.1 Coding by the respondent

This solution will in practice mean that the respondent is requested to place his/her job in one of a set of pre-defined groups presented to him/her in written form on a questionnaire or on a card handed over by the interviewer/enumerator. The main advantage of this approach is that it is the least expensive of the possible procedures. The main disadvantage is that it results in low quality data, both in terms of reliability and validity. The low level of reliability is due to the difficulty in assuring consistency in how the respondents relate the pre-defined groups to the job they are asked to give information about. The content of each group has to be described in a limited number of words, normally in the form of a group title, and it may be very difficult to give respondents the intended understanding of how their job relates to the different groups. The fact that the number of possible groups to choose between is limited will of course limit the possible number of mistakes to be made, however, studies have shown that when coders code to a hierarchical classification, many of the errors made are in the choice between the limited number of aggregate groups. The limited validity and usefulness of the resulting data is a direct consequence of the limited number of groups to which coding can be done when using this approach. Many users of occupational statistics and information in general need to make much more detailed distinctions and to work with much more homogeneous groups than those obtainable when using this approach. However, the cost advantages are such that some national statistical agencies and most commercial survey organizations have been using this strategy. Administrative agencies may also choose to use this strategy, in particular if the need of the agency is to establish whether or not its "clients" can be classified to a very limited number of specific occupations, or only to sort them into a few broad groups.

4.2.2.2 Coding by field staff

With this strategy one possibility is that the interviewer/enumerator assigns the response to a pre-coded alternative 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 interviewer/enumerator to understand and 'translate' the

information received to the appropriate group. This procedure should be used if the respondents themselves are not expected to read the questions and write the answers. The consequence of this procedure for the validity and usefulness of the resulting statistics is as described above. The reliability of the coding may, however, improve compared to coding done by the respondent because the interviewer/enumerator may receive much more detailed instructions on what types of jobs the different pre-coded groups are supposed to cover and where there are ambiguities which will require probing. On the other hand the interviewer/enumerator may misunderstand what the respondent tells him and therefore not select the correct group.

The other possibility for coding by the interviewer is that he/she writes down key words of the respondent's answer and then code the response after the interview, but before the questionnaire or is forwarded to the processing centre. The advantage of this procedure over those mentioned above is the possibility this gives for much more detailed coding and thus much more valid and useful results. An interviewer coding outside the immediate interview situation may be given complete coding indexes to assist the coding process, as well as other coding aids - including the possibility to forward queries to supervisors. An advantage is, however, that the interviewers will often retain in their memory more details about the response than they are able to write down on paper, and therefore have a better basis for selecting the correct code than specialized coders who have to rely on what has been written down. Another advantage is that as the interviewers gain experience with coding they will become more aware of the type of information which is required to code correctly. Compared to the use of specialized coders described below, the main disadvantage is that because they are more numerous and often geographically more scattered, the interviewers cannot be given the same amount of training, supervision and support as specialized coders, with consequent results for coding reliability.

A statistical consideration with this approach is that all individual coders have irreducible idiosyncrasies which produce biases, relative to the mean of all coders, in the distribution of codes which they allocate. It can be shown that, if biases of similar magnitude are imparted by all coders, whether in the field or in the office, they will have less effect on the overall variance of results if each of a large number of field workers codes a relatively small number of cases, than if each of a relatively small number of office coders codes a large number of cases. Depending on the relative sizes of the work quotas of field and office workers, the variance of estimates may still be lower even if field workers individually perform as coders in a more variable (less consistent) way than office coders.

It may be possible to save time, costs and complication by resorting to field coding in circumstances where occupational data is the only item requiring office coding in survey documents which could otherwise be passed straight on to computer data entry. On the other hand, field workers may be thought unsuitable in terms of background and training to act as coders. Also, field coding sacrifices to a large extent the important advantages of a controlled and supervised coding environment.

Field coding by interviewers can be compared to the situation where coding is carried out by the staff of local offices of e.g. the employment service, where the same advantages and disadvantages as those outlined above will apply. However, an added consideration with respect to the administrative staff is that in their case the consequences of poor coding quality may be experienced by themselves or by some of their close colleagues, thus providing them with a direct feedback on coding quality and an incentive for correct coding.

4.2.2.3 Office coding

Most national statistical agencies use specialized, centrally located coders for the coding of occupational information for their population census and surveys. Such coders may be entirely specialized on occupational coding, or on coding in general lie. of industry, occupation and sometimes education and training received (qualification)] or they may carry out occupational coding as one part of an integrated data entry, coding and data control operation. The exact context will depend on the whole organization of the census or survey processing operation, which again will depend on the scale of the operation and whether it is continuos or ad hoc. As indicated above the advantage over the use of interviewers as coders is the better possibilities for training, supervision and immediate query resolution, while the disadvantage is the increased distance to the original response and the complete dependence on what the interviewer/enumerator or the respondent has noted down on the questionnaire/schedule.

Where the collection of occupational information takes place on a continuous or regular periodic basis, coding is likely to be done by a unit which is permanent and not very large, employing staff over long periods. In such circumstances staff will become thoroughly accustomed to the practical routine of the task, the rates of work required and so on. However, there is a tendency for such units to become strongly 'production minded' and cut off from the rest of the organization, with little incentive to assess their product in terms of external validation criteria. After a time few new queries or operational difficulties are seen to arise and managers and data users often take this as evidence that the task presents few problems and is being carried out to high standards of accuracy, reliability and validity. Such an assumption and situation is, however, dangerous. It is probable that intractable coding problems have not actually been solved, but rather dealt with by ad hoc rulings. Also, individual coders tend to identify 'short-cut' methods which reduce the laboriousnes of their task but which may also incorporate errors or unjustified assumptions, if not actual violations of the coding instructions. These unofficial departures from the 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 system of classification in use. More generally, issues of external validity and internal consistency of coding tend to fall into abeyance, unless the routine procedures of the unit include specific and properly designed checks on the levels of coding validity and reliability achieved. The use of specialized coding units should therefore be avoided in on-going operations. Coding of occupation should be one element of a larger data and more varied processing task.

The large-scale occupational coding exercises mounted on national censuses of population, for example, generally rely on special recruitment and training of inexperienced coding staff. For reasons discussed above, the performance of staff who are inexperienced, but suitable and adequately trained and supervised, may actually be superior to that of staff who over a long period have acquired bad coding habits. Nevertheless, the results are likely to be poor unless recruitment, training and production 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 census cycle, there may be insufficient time to establish rules and routines and generally 'run in' the coding establishment before production coding has to start. In such a situation the organization may 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.

4.2.2.4 Concluding remarks on field or office coding

The arguments and counter-arguments outlined in this section must be weighed in order to decide whether field or office coding of occupation should be used in a given context. Ideally, of course, objective data on the performance of field and office coding are needed. Several studies have been undertaken in the UK to obtain such comparative data:

Interviewers on the permanent field staff first received special training and instructions in carrying out occupational coding. In the field they then asked standard questions on job title and job activities, probing further in cases where they felt the initial answer to be inadequate and recording the responses verbatim. They recorded the responses in full and coded them outside the interview situation, using an occupational coding frame containing around 350 categories. The verbatim material was then re-coded independently by office coders and the codes allocated by field and office coders were compared. It was found that average agreement rates were of the order of 70 per cent which compares with an average agreement rate between office coders of 85-90 percent. In interpreting this finding account should be taken of the results discussed earlier on the validity of the field data collection/office coding methodology. These suggested that, when data are obtained for the same individuals by two independent approaches, the average agreement rates for occupational codes allocated may only be of the order 70-75 per cent.

The general implication is that, in censuses and surveys which rely on simple questions to establish occupation, the chances that an occupation will be allocated to the correct one of several hundred occupational groups may not be more than seven in ten. Trained field workers probably do not code as consistently as trained and supervised office staff, but it is by no means clear that the overall balance of advantage is in favour of office coding. However, the arguments in favour of using field coding, as compared to specialized coders detached from the data collection process, are closely linked to having a permanent field staff which cost effectively can be trained also in occupational coding and which can accumulate experience. This means that the field staff turnover rate should not be higher than for the specialized coders. It also means that field coding is a realistic option for continuos labour force and similar surveys, as well as for the registration of "occupation" by local administrative officers, but not for ad hoc and much larger scale operations like population censuses.

4.2.3 Level of coding

The purpose of the coding process is to determine and record, from the responses obtained from the respondent, to which of the groups in the classification the job of the respondent belongs, while retaining as much as possible of the information contained in the response. The coding process can be seen as a translation process where the coder 'translates' the occupational response into the correct Occupational group, because only in a minority of cases will the words of the response be the same as those used to designate one of the occupational groups. For the coders the raw materials for this process are the responses and the tools are the coding index, the coding instructions and the persons responsible for answering queries.

The more information retained after the coding, the more valid and therefore valuable are the occupational statistics for the users. The coding process should therefore be designed to find and record the most detailed code supported by the response. This is contrary to the praxis of many statistical agencies. Their most common procedure has been to decide that coding should be done at a particular level in the classification, e.g. the 3 digit (unit group)

level. The arguments for this have commonly been: a) that it would be too costly to code to a larger number of groups, both in terms of coding errors and in terms of working hours; b) that the responses would not support coding to more detailed groups; and c) that (in sample surveys) 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 country experiences 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. At any rate, the error rate for more aggregate groups does not seem to increase, on the contrary. One country has estimated that to code at the occupation level instead of the unit group level, ie. to increase the number of groups from 280 to 1 100, would require an increase in the coding index of about 5 percent.

b. Experience clearly shows that the occupational responses are very uneven in the level of detail they will support for the coding. Many responses will support detailed coding, especially if the questions are formulated along the lines outlined above. A significant number of responses will, however, not even support the level conventionally chosen. By insisting on a pre-defined level, the coding process may both lead to unnecessary loss of information for a large part of the returns and to misrepresentation of the data quality for other parts.

c. Similarity in type of work performed is the general criterion used for defining the groups in the occupational classification. This means that the number of jobs which can be found in groups defined at the same level in the classification, will be very differ greatly, even if 'statistical balance' has been one of the supplementary criteria used in the construction of the classification structure. The number of jobs in a high level group may therefore also be smaller than that of a detailed level group. In addition, the tabulation of occupational data also typically involves both the merging of groups and cross-classification with other variables such as age, sex and industry. Consequently one should not unnecessarily restrict tabulation possibilities during the coding process.

4.2.4 Coding of vague and difficult responses

Most occupational classifications specify residual groups of “type x occupations not elsewhere classified" - usually as separate unit groups within some of the minor groups or as occupational groups within unit groups. These groups are designed to take care of jobs and occupations that belong to the more aggregate group but which are not similar enough to any of the specified sub-groups within it to belong in either of them and which in themselves are too small in numbers to warrant separate specified groups. This means that they should not be used to code those responses which the coders cannot assign to any of the specified groups. Such responses can be either: (a) too vague and imprecise to allow the coder determine which group the job belongs to; (b) indicate that the job in question has tasks and duties which cut across the distinctions made in the classification; or (c) represent a type of work not covered by the classification.

The proper way of handling such 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 belong. 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 group together with those jobs which properly belonged to that group.

b. The classification of jobs with an uncommon mix of 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 left to expert coders or raised as queries for 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 jobs with tasks and duties not covered by the classification. The reporting of difficult cases of these types is an important input to the process of updating, maintaining and possibly expanding and revising the classification.

The basic information to be used as a basis for the coding of occupation is that contained in the responses to the occupational questions outlined above. However, in a significant number of cases where the occupational information is not sufficient for the coder to choose between possible alternatives, ancillary information provided by the responses given to other questions may provide the basis for making the choice. The most important type of such information is that pertaining to the 'industry' of the workplace. In addition information about educational and vocational qualifications may also sometimes be used. It must be emphasized that the use of such ancillary information should be strictly and specifically controlled in order to avoid an undermining of the descriptive and analytical use of 'occupation' together with either 'industry' or 'qualification'. This means that rules for the proper use of such ancillary information, i.e. about when and how to use it, must be incorporated into the coding index and the coding instructions.

4.3 Planning and organizing coding operations

The discussion which follows will be based on the assumption that coding is to be carried out by a specialist coding staff in the context of the final processing of the census or survey. This is because this is the solution most commonly chosen by national statistical agencies, and because it is, in most cases, easy to see whether or not similar guidelines apply to field staff coding.

Assembling the right resources for processing of occupational data in the right places 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 coordinated with other processing tasks. Large volumes of documents and data have to be handled and, because of the interdependence of different stages within the overall processing plan, the penalties, in terms of delays and cost increases, for failures of operational or quality controls may be heavy. The main outlines of the processing plan for a population census, particularly as they affect requirements for finance, staff, equipment and premises, may need to be worked out long before the start of the actual processing. Managerial staff who will be involved in the planning and supervision of the coding operation and professional staff concerned with the design of the classification and coding procedures, the training of coders, the updating of classifications and interpretation of results need to collaborate closely at the

planning stage.

4.3.1 Finance and resources

Substantial amounts of money are required to support the processing of a census or a large scale survey 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 adequate provision. 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.

4.3.2 Expertise, experience and rehearsal

The coding of occupational data demands special expertise in those who plan, manage and supervise it. Processing of each census or survey relies heavily on technical information, expertise and experience gathered at the last similar exercise. Information and experience should be documented in detail, but practical expertise is likely to reside very largely 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 adequate to rely entirely on documented or undocumented experience from the past. 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 costs and time requirements, based on concrete experience. 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 it is still possible to adjust plans for the main operation in the light of them.

4.3.3 Estimating coding rates

Some key planning parameters (e.g. coder work rates, effective document throughput rates) can only be reliably estimated from previous experience or planned tests. 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. Experience also 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. There is, however, the danger that this effect 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.

4.3.4 Coding staff

It is important to make good estimates, not only of the number of coders required, but also of the numbers of first-line supervisors needed to control the coding process and the number of specially trained staff needed to resolve queries.

In a census operation the large volume of work to be processed within a limited period will require special staff recruitment and training, both for initial recruitment and for anticipated staff turnover in the course of the task. Thorough inquiries 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. Common criteria should be defined and applied 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 at both the coder and supervisor levels, and to recognize that the specialists normally cannot, and should not, be recruited and trained for a single survey or census operation, but be part of the permanent competence of the organization.

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

a. Literate and reasonably intelligent, but not over-eager to display independence of judgment 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 coding staff 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 task). Mistakes at the initial recruiting stage are likely to lead to high staff turnover, and the need to recruit and train replacements while production coding is in progress.

Occupational coding does not require theoretical knowledge. It is best learned through practical instruction in specific procedures (e.g. document handling routines, use of the coding index etc.), interspersed with supervised practice on appropriate, specially designed exercises. It is important to identify at this stage persons who are unwilling or unable to follow instructions. Relative slowness in learning need not be a disadvantage if accompanied by good retention of what is learned and the desirable temperamental characteristics.

4.3.5 Coding teams and supervisors

Production coding on large jobs 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. First line 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 occupational coding 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 occupational 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 occupational classification and coding system.

The number of coders allocated to each supervisor is important. Typical ratios lie between 6 to 1 and 12 to 1, but the appropriate ratio 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 is likely to cause not only bottlenecks and poor staff morale, but also 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 where coders are persons who 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.

4.3.6 Coding tools

It will be necessary to provide appropriate documentation, procedures and training material not only for coders, to guide the coding process itself, but also for supervisors. The basic tools required by coders will include:

a. **Coding instructions**: These should cover all operations which the coder is required to carry out. It is likely that the same coders will deal with several different data items besides 'occupation'. The procedures and instructions for handling all relevant items and operations should be integrated to take account of this. The instructions relating to the coding of occupation will need to be particularly clear and specific on:

   - the order in which checking, coding and editing tasks are to be carried out; the procedure for analyzing verbatim material for significant terms;
   - the use of the coding index;
   - the circumstances and procedure for using ancillary data.

b. **Coding index**: This is the key coding document through which verbatim terms incorporated in job titles, descriptions of tasks etc. are translated into codes. Coders should not be encouraged to interpret verbatim responses in terms of their own conception of the purpose or criteria of classification, but rather to follow in a conscientious way the instructions laid down for consulting the index. For these reasons it is obviously essential that the index be clearly set out, explicit and easy for coders to use (see chapter 4). The use of the coding index, instructions and procedures needs 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.

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 classification itself.

d. Legal and administrative forms: These may include a legally binding undertaking, to be signed by coders, to maintain the confidentiality of census data. Other documents used by both coders and supervisors are likely to include forms for recording queries and their resolution; for controlling the flow of work and reporting progress; 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—may be—should be used, e.g. posting of productivity and error rates for coding teams.

4.3.7 Coding problems and queries

No matter how carefully coding instructions and the coding index have been prepared, it can be guaranteed that large numbers of detailed queries will be raised in the course of a major occupational coding operation. This happens mostly because the index can 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 most carefully designed questions and instructions to interviewers/enumerators. Any revision of the structure of the occupational classification since the last census or survey may also lead to a new crop of problems in the treatment of vague and inadequate responses at the borderlines between categories.

Evidence of shortcomings of the documentation thrown up in the course of production coding will need to be rapidly and consistently processed and fed back in the form of amendments. 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 in the coding documentation and procedures. The roles of supervisors in processing queries and amendments need to be defined. Particular care is needed in the coordination of query reporting and document amendment where coding is being carried out in several different locations, e.g. in the different local offices of the employment service.

4.3.8 Quality assessment and quality control

Casual observation by supervisors and crude visual checking of coded output do not provide adequate information on quality of output. Explicit allowances for the resource and time costs of formal quality control need to be built into the processing plan. These will cover the establishment and staffing of a quality control unit responsible for acceptance testing of coding as the operation proceeds and for assessment of the reliability and consistency of the operation as a whole. A procedure for sub- sampling the work of occupational coders for quality control purposes needs to be defined and 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 must make allowance for corrective action (e.g. 100 per cent checking) in the case of batches which fail a quality control test.

In-built quality control procedures will also be required. It is necessary to design separate procedures to handle:

(a) on-line acceptance testing of coders' work; and

(b) overall monitoring and assessment of performance.

The aim of acceptance testing is to identify rapidly coders whose performance does not meet criterion levels of accuracy in following coding instructions, 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 occupational 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 cases. On the basis of such quality assessments it may be possible to separate the contributions made to total error variance by error/variability in data collection and error/variability in coding.

4.3.9 Premises, infrastructure and equipment

The large volume of census and large scale survey 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 bearing personal details. Proper attention must also be made to the fact that coding is a monotonous task which makes staff sensitive to the work environment, the functionality and capacity of the desks, chairs, shelves, filing cabinets heating and ventilation as well as the adequacy 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 are to operate with special equipment, e.g. computer terminals, special arrangements may be needed to estimate the requirement, identify suitable and reliable equipment and carry out tests, estimate and provide for capital expenditure and depreciation, provide for replacement in case of breakdowns, go though procurement procedures, etc.

4.3.10 Process in one location or several?

Census and ad hoc survey processing creates a substantial but temporary demand for suitable staff and premises and this may cause extra staff recruitment and other management problems if all processing is carried out in a single location. There may also be other cost and logistical arguments for carrying out processing at one or several locations other than, or in addition to, the central census processing office. On the other hand it needs to be borne in mind that, for relatively complex tasks such as occupational coding, it is both difficult and important to maintain team consistency in coding between coders and coding teams. One reason for this is that occupational coding inevitably generates large numbers of queries (in the 1981 UK Census more than 30,000 coding queries were processed and it is thought that many more were dealt with informally). These in turn lead in some cases to amplifications or changes to the coding index and instructions, which then need to be applied in a consistent fashion. This, and the maintenance of consistent production and quality control standards more generally, is more difficult to achieve when coding is carried on in several locations than if they are centralized to one place.

4.3.11 Handling of documents

The coding of occupation will normally be an integrated part of the total processing of the information on the administrative forms or the census/survey questionnaires. In that case the main concern will be:

- how to receive the form;
- how to store them; and
- how to allocate them to staff so that one can control that forms have been processed.

It should be feasible to find individual forms which for some reason need to be

rechecked. If each form 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 loss of forms.

4.3.12 Use of computer assisted coding

A basic practical choice which affects the logistics of data processing is whether to write the codes on to the data collection forms, which then function as computer data input documents, or to integrate the coding into the data entry process, either by keying them directly into a computer file after they have been found from the coding index, or by having them (semi-)automatically assigned through some *computer assisted coding (CA CJ* system. If it is intended to introduce a CAC 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 reversion to a manual/clerical system as a fall-back position.

5 The development of coding indexes

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 have mainly originated in such countries. Unfortunately, our knowledge of the experience with this type of work in other languages is limited and it is therefore difficult to say to what extent the documented experience is transferable to other languages and cultures. This should be kept in mind when reading this section. However, the experience from English- speaking countries may provide a starting point for work and experiments in other languages if nothing more appropriate can be found.

5.1 Defining the index

The process of coding occupational information involves the task of matching responses against index entries, to derive an occupational code. The coding index is the key instrument for this matching process. The index can take the physical form of a durable printed publication, a loose-leaf binder, a computer print-out or a machine-readable file within a computer system, and the matching can be carried out by a person, i.e. the coder, by a computer or through the interaction between the coder and a computer.

Most detailed occupational coding is still carried out using clerical procedures. Job titles and descriptions are recorded verbatim in the field by members of the public, enumerators or interviewers and the resulting raw data are brought to one or more central offices. Here clerical staff (coders) scrutinize each case, decide (suitably guided by coding index and instructions) to which occupational unit group to allocate it and record an appropriate code on to a document, or directly on to a computer-readable medium for further processing.

The coding index is the principal instrument used to link the words used in the various parts of the response—job title and job description —to the numerical form which represents the allocation of that job description to its position in one of the groups of the classification. The coding index guides the coder by listing key words which can be found in the responses to occupational and other ancillary questions and indicates how different responses are allocated to the detailed or more aggregate occupational groups, the 'building-blocks' of the classification, depending on the nature of the response and the instructions for the coding process.

It is important to recognize that a coding index in principle is different from the other two indexes which are often associated with an occupational classification and dictionary, namely the index of occupational groups defined in the classification, and the detailed index of occupational titles. The former is just a list of those occupational groups which are separately defined in the classification and dictionary. The titles chosen for these 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 their jobs. The latter index consist of titles which also are chosen to be descriptive of the content of the groups, i.e. to illustrate their content, and may therefore contain entries which may never be used as a title by any person describing Their job. When they exist both types of indexes may, however, serve as useful starting points for the construction of a coding index.

5.2 Developing and updating the coding index

The first issue relating to the development of a coding index concerns the nature of the index itself. Basically the choice is between two different approaches. Some countries have adopted the approach that the index should be all-inclusive: i.e. every distinct occupational response found in the process of occupational coding should, in theory, have an entry in the index, allowance 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, when faced with an obscure job title, to find that title 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 verbatim indexes create the impression that occupational 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 job descriptions fail to match the index entries, and one has to use rules and/or judgment to make the 'best' match. For these reasons, other countries have developed a structured index.

A structured index does not try to reflect every possible occupational response because it is combined 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 identify the occupational group, an appropriate qualifying word (or phrase) must be added to distinguish between the possible alternatives. 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 a response into an entry in a structured coding index according to the following format:

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

Examples:

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

The following examples from the coding index used for 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 - ASCO):

5999 Researcher/market/interviewing
2909 Researcher/market/statistician
2907 Researcher/market (except above)
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 Signwriter
4921 Silverer/glass
4923 Silversmith
2815 Sinner.

The much smaller number of index entries in a structured coding index than in a complete listing index is a result both of the restriction of the index to functional words when possible and the use of '(except above)' instructions.

The functional word is the generic part of the job title found in the occupational response, i.e. the word which standing alone can serve as an occupational title, while the qualifying words usually indicate some form of specialization. Sometimes the functional word may be precise and in itself suffice as an index entry, cf. 'Signwriter' in the examples above. However, the functional word may also be very ambiguous, cf. the '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.

The construction of the structured coding index must reflect and support the coding rules to be used for assigning the occupational code on the basis of an occupational response and the permissible ancillary information given in other responses. 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 permissible parts of the responses, and the qualifying words must reflect those which can permissibly be selected, in the order of priority in which they should be selected.

In English it will be normal to use as functional words that can be found in either (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 title or task components of the response. Second priority, and based on the rules for using industry

information in the coding of occupation, should be given to words found in the industry or name/type of employer response. In the examples above the qualifying words which are underlined signify that the coder is allowed to use the industry or, if that is not sufficient, name/type of employer response.

The advantages of creating a structured coding index are twofold. First, it causes 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 coder's search through the index because of the smaller number of entries.

Some words may be found in common usage in job titles, but can be ignored for the purpose of creating the 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.

Given the resources which must go into the construction of a good coding index, it is usual to design an index to last for a significant period, say 10 years, and for the index to have wide usage in various government departments, by survey agencies and academic users. If this is the case, a well-bound quality publication may be considered. A different strategy may be to design the index for regular updating, in which case a ring-file or computer print-out may be the appropriate format. In many instances it may prove appropriate to consider a combination of these two alternatives.

The physical form of the index and the method of updating the index go hand in hand. One of the major problems relating to most index systems is the fact that they are often out of date soon after publication. If they are based upon information contained in a census and the census is held on a decennial basis, the index may not reflect changes in occupational terminology and structure in recent years. Some method of keeping an index updated on a regular basis must be sought.

The process of updating an existing index should be viewed as part of the general processes required to maintain an occupational classification. As new ways of organizing work between or within enterprises or new technologies are introduced, new jobs will appear with new combinations of tasks or new types of tasks associated with them. These new jobs may be given new job titles, 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 being changed in any significant manner, e.g. as a result of reorganization of the enterprise or because their placement in a wage hierarchy has been changed. Thus, there is a need to keep track of job titles and the associated job descriptions, monitoring the relationship between this information, index entries and the associated occupational codes. A full scale job content monitoring exercise cannot be used for this as that would be too time-consuming and costly. The most realistic alternatives would be the following two procedures:

(a) post-coding reviews

(b) job vacancy reviews

The former of these procedures uses the coding team to assist with the review of the index. Coders are a good source of information on the adequacy of an index. Their suggestions for improving the 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 index material should be designed into the data processing operation. It is essential that information which may be useful in updating classification and index be carefully collated

and retained. Such information typically consists of records of problems encountered, queries raised and decisions and amendments to the working instructions adopted in the course of coding. From the point of view of a Population Census operation the disadvantage of this approach is that the index reaches its most up-to-date state at the time when it is least likely to be used, immediately after coding is complete. Nevertheless, when the index is used for a variety of registrations, and in particular for continuing surveys and administrative operations, the use of a post-coding index review procedure must be given serious consideration. Such procedures should of course be made part of the normal routine for continuos or 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.

When available from job advertisements in newspapers, journals and/or registrations at a local office of the employment services, job vacancy descriptions may provide a useful source of up-to-date information on job titles and detailed job descriptions - especially where these have been coded to the occupational classification as part of the job vacancy recording process. By sampling such coded job vacancies on a periodical basis, it is 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 updating the index (and the classification) accordingly. The advantage of this approach is that it does not require expensive initial search for contacts data collection, as follow-up inquiries to an interesting vacancy will have the name, address and contact person of the employer directly from the vacancy notice. The main disadvantage is its reliance upon job vacancies which have been advertised or notified to and recorded at an employment agency, as these recruitment channels, normally only cover job vacancies for a limited range of occupations and industries.

Some countries have established standard procedures for the collection of job vacancy materials. 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 be used within a word-processing system or may form part of a database of vacancy information. In such cases, it may be possible to extract this material and incorporate it with the coding index and classification. This technique can be performed on a regular basis, maintaining the index as an up-to-date record of job titles in current usage. Similar procedures can be used when there is an effort to update occupational descriptions developed to provide information for the planning of vocational training courses and/or the material produced for vocational guidance.

5.3 Using the coding index

5.3.1 Using the occupational response

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 appropriate code in the occupational classification structure. The main tools for this 'translation' are the coding index 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 how this translation process should be carried out; what items to look for in the occupational response and in what order; what type of ancillary information to use from other responses; when such information can permissibly be used and how to use it. Ideally the coding index should have been constructed to reflect these instructions.

The starting point should always be the response given to the occupational question(s), i.e. the question(s) asking for the type of work in the person's own job and the

usual or main tasks and duties carried out in the job. In English this question should normally result in a job title and a few words on main tasks. The coder should start by identifying the job title part of the response and start looking in the index for this title. 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 'school cleaner'. 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 aIternatives:

(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.

5.3.2 Using ancillary information on industry or name and type of employer

Most modern systems of occupational classification, including ISCO-88, are designed on the principle that 'occupation', meaning a particular pattern of work tasks and skills which constitute an individual's job, should be kept conceptually separate from 'industry', meaning the sector of the economy 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 the coal mining industry 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 description. For example, the occupational term 'coil winder', used on its own, is ambiguous because coding depends on whether what is wound is some form of metal wire, some form of textile product etc. Knowledge that the job is located in the textile manufacturing industry may be sufficient to resolve the ambiguity with a reasonable degree of certainty.

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 has to consider extra sources of information 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 information in different ways. These latter two problems are minimized:

(a) if industry is coded in advance of occupation, so that no further interpretation of verbatim responses is required of the occupational coder;

(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 an occupational coder can make on the basis of the industry code are exhaustively pre-specified through index-referenced instructions.

A simplified example may clarify this: A coder encountering the job title 'coil winder' would be instructed in the coding index under 'coil winder' 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.

In practice occupational coding has tended to rely quite heavily on industry information. For example, an experiment using data and coding procedures drawn from the 1981 UK Census suggested that the occupational code allocated would have differed in about 19 per cent of cases if no use had been made of pre-allocated industry codes. A dilemma which may arise in these circumstances is that, whereas use of industry data in a particular context may significantly improve the quality of the occupational coding, by the same token it must reduce comparability with other sources where coding must be done without the benefit of data on industry. In such circumstances coding will be carried out under two different sets of coding instructions and the comparability conferred by the use of the same system of classification may be seriously compromised, assuming that the absence of industry information would lead to the same, less sure and less precise occupation codes, in the two data collections. This underlines the need to look at the variables and questions related to respondents' economic activity in a coordinated and integrated way, based on a concrete knowledge of both how the occupational classification is constructed and on the division of work in the labour market, when planning and executing a population census or survey, or in the administrative records. If the use of the industry information is not controlled, but left to the judgment of the coders and their supervisors, then the coding process may introduce into the resulting statistics a pattern of occupations across industries which does not exist in reality.

5.3.3 Use of other ancillary information

Some countries 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 corresponding 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 training, e.g. because he/she lack necessary language skills.) Therefore, the use of information on qualifications as ancillary information should therefore be very carefully controlled and probably restricted to query resolution by expert coders.

5.3.4 Inadequate responses and queries

Some responses simply cannot be coded to a detailed occupational group. This will normally be for one of the following reasons: (i) the response may be vague, i.e. not contain enough information to be coded according to the coding index and coding rules; or (ii) 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.

Unfortunately, the number of cases of type (i) is likely to be quite substantial, even with well-formulated occupational questions and well-trained and experienced interviewers. 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 used in the response, cf. the use of '1 503 Restaurateur' and '5101 Secretary' as default groups. 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 '2000 Researcher' indicates that the response could only be coded validly to 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, which is not likely.

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 first line 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 first line coders and a very large query burden on the supervisors. It may therefore be better to monitor carefully the use of 'default' codes by first line coders.

The fully specified responses which are not adequately covered by the classification 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 occupations “not adequately covered" by the classification.

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 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 required, as

in emergencies, and because time allocation of tasks is an information normally not available.

Precise responses which cannot be resolved by priority rules should be placed in special “not adequately covered" or “general tasks" groups created for coding purposes within the aggregate groups to which the jobs clearly belong. Steps should also be taken to ensure that these cases can be closely examined outside the coding operation itself to determine what, if any, contribution such cases can make to the updating of the coding index and of the classification itself. Note that these groups are not the same as the 'not elsewhere classified' groups of the classification. Great care must be taken not to make a confusion between the two types of groups.

6 Use of computers in data collection and processing

6.1 Background

Up to the mid-1980s, most countries would have reserved computer usage for what are termed downstream applications. That is, after data had been collected, manually transposed to data collection sheets and key punched on to cards or paper tape or entered on magnetic tapes or disks, the machine-readable data would be fed into a computer (usually a large main-frame computer) and verified through a series of computer edit checks. This process would involve passing the data through a computer program to ensure that any out-of-range codes or non-allowable combinations of codes were detected and reported for further investigation. These types of controls are still valid and important, but, computer edit-checking cannot detect invalid coding in the range of valid codes or if errors in data collection and coding do not generate non-allowable combinations of data. Apart from edit-checking, data processing by computers was usually reserved for the final stage of the whole operation, the production of statistical tables.

More recently, data processing has benefited from the development of database management systems. For large-scale data processing applications there are main-frame versions of these systems which allow the non-specialist to design screens and forms and to develop a system for the processing of data at all stages from data entry, through edit-checking, management of the processing facilities to the presentation of statistics for publication. Previously, the manipulation of census information or large-scale social survey information via such systems would have been prohibitively expensive in terms of the hardware requirement for on-line data storage. Technical advances in recent years have now made large-scale database management a feasible proposition for such applications.

6.2 Data collection

The most important development in this area is the use of portable microcomputers for the collection and coding of occupational information on a decentralized basis. In an 'idealized' version of this system, the field worker or interviewer contacts the respondents by telephone or collects census returns or survey schedules from households and, working from home or a decentralized office, runs a database management program on the microcomputer to assist with data entry, automatic encoding of data entries, data verification and compilation or extension of the machine-readable database. After sufficient entries have been stored within the microcomputer by its computer program, either on floppy disk or in solid state memory, the data are then transmitted in machine readable format, possibly via the telephone network, to a central processing computer. This process continues until the central computer recognizes that all of the decentralized locations have reported, at which point the data are immediately available for processing.

Such systems have many advantages over traditional data collection and processing methods. Higher levels of standardization can be achieved, the interviewer may check back with the informant if inconsistencies are detected in the data and the length of time from collection to publication of statistics can be greatly reduced. Also, such a system can generate much useful information for management purposes. For example, when used by specialized coders or data entry staff, the rate of entry of data into the microcomputer can be recorded automatically, leading to accurate estimates of occupational coding rates.

6.3 Data reading and coding

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 occupational coders. Recently, however, the situation has begun to change rapidly and systems for automatic or computer-assisted coding of occupations are now used in a number of countries. The introduction of these methods may well have had beneficial effects on the consistency of coding, but they do not reduce very dramatically the combined coder and data entry working time and the elapsed time required to complete the task unless it is possible to use the equipment in a way which will significantly reduce the task of transcribing the verbatim material into a computer-readable form. Even where operators with appropriate keyboard skills are available, this operation is inevitably slow and its role as a limiting factor becomes more obvious as the power, efficiency and speed of the computer hardware and software increases. Some countries have reported that the use of optical reading forms and equipment contributed significantly to the processing of their population census in the 1990 round, and in some cases this provided the basis for the introduction of computer assisted coding (CAC) systems for occupation, industry and educational attainment (qualification).

So far no workable system has been developed which fully automates the coder s decision-making task. Some 'automated' coding systems claim to allocate codes automatically to more than 70 per cent of cases, but their development costs have often been high and there have been problems in making them sufficiently 'intelligent' to simulate reliably the performance of trained human coders. The reported error rates for those responses which the systems code are 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.

6.4 Computer-assisted data collection

As mentioned above one promising area of development is that of 'lap-top' computers and appropriate software offering computer-assisted coding of occupation as one element in a computer-assisted interviewing package where the office coder is replaced by a field interviewer interacting directly with an informant, while being guided by a computer-based interview schedule and coding instructions. Computer assisted interview packages were initially developed for application to office-based telephone interviewing. The increasing power of portable microcomputers has recently made it feasible to deploy similar packages in field interviewing also. Such packages control the interview by displaying the appropriate questions and response categories on a screen. They automatically handle continuity and dependencies between questions and carry out certain edit-checks to ensure that the responses input satisfy format and consistency rules. Responses are entered directly into the computer via a keyboard and the data recorded in the course of interviews are periodically unloaded into a larger computer system which carries out further data manipulation and data analysis. Many packages include facilities for prompting interviewers and (in the face-to-face situation) informants with lists of codes and the like.

For this application the microcomputer's memory and data storage capacity needs to be sufficient to store: the programs which control the interview; the lists and coding frames appropriate to particular data items; and a reasonable number of interview data sets. This requirement to store lists of codes and to display them in a way which lends itself to accurate coding is, of course, a demanding one in the case of occupational coding but portable

microcomputers are now reaching the levels of power, speed and memory capacity required. Progress may well be speeded up by assimilating work which has in recent years been done in different organizations, some working on Computer Assisted Data Collection systems and others on Computer Assisted Coding systems.

The attractions of such an integrated, computer-assisted method of capturing and coding occupational data are the following:

(a) The data are entered directly on to computer at the point of capture, thus cutting out the time-consuming and potentially error-prone data transcription stage referred to above.

(b) The computer is able to check rapidly whether the initial input is adequate to identify uniquely the occupational group to which a job should be allocated and, where necessary, to issue prompts asking for appropriate further information.

(c) All this takes place while the interviewer is in direct contact with the informant and can put appropriate supplementary questions to elicit any additional information required for coding purposes. This is likely both to enhance the validity of the code recorded and to eliminate to a large extent the need in classifications for 'dustbin' categories to which ill-defined occupations can be allocated.

The method thus combines the strong points of the human operator (intelligent flexibility and ability to interact with another human in appropriate ways) with those of the computer (speed and accuracy in handling large, diffuse but logically defined tasks, such as searching indexes, and consistency within a pre-defined range of operations and decisions). It avoids the respective weaknesses of the human (slowness and error-proneness in carrying out routine but complex tasks) and of the computer (inflexibility in handling situations which have not been exactly anticipated and provided for).

On the other hand, an obvious limitation is that the approach requires the presence of a field worker interviewer armed with a microcomputer and is thus not suitable for data collection on the scale and in the circumstances of a census. Even for more limited surveys, special consideration must be given to the use of such a system which must be safeguarded against technical and/or human failures. Computer data recording and storage media are much more vulnerable to damage than pen and paper. Although the risk of damaging large sections of data is low due to the decentralized collection procedure, the risk of any damage occurring may be higher than with traditional methods. Also, consideration must also be given to the need to maintain data (and microcomputers) in a secure environment. A file of census or survey forms may not prove an attractive item to a thief, but the same cannot be said of a microcomputer. Thus, the design of a decentralized and computerized system of data collection and initial processing must have safeguards built in to it to minimize these risks and to facilitate the recovery of data from damaged storage media.

6.5 The experience of the Australian Bureau of Statistics

The Australian Bureau of Statistics (ABS) introduced a new occupational classification for official statistical purposes just before the 1986 Population Census. This classification differs significantly from its earlier classification through its use of the skill level of an occupation as a distinguishing component and the aggregation of occupations by skill levels and skill specializations, as in ISCO-88. The introduction of this completely new classification yielded the opportunity to develop new procedures and instructions for the

coding of occupational information, designed to ensure even and drastically improved quality in the resulting occupational statistics. In order to achieve this without an equally drastic reduction in coding rates the ABS developed a computer assisted coding system which was applied to the coding of occupational information for 100 per cent of cases in the 1986 Australian Census of Population and which has later been further developed:

The new classification required the coder to use all of the available information from the job title, associated description of tasks performed, name of employer and nature of employer's business. This information was used by the coder to obtain a match to the nearest entry on a 'tree-like' computerized display of the index. The complete system is very elaborate, but a simple example will indicate the logic of the approach. The responses to census questions on title, tasks, employer's name and type of activity (industry) may be as follows:

Title: Insurance Office Branch Manager
Tasks: Running the Branch
Employer: Govt Insurance Office, Canberra, ACT
Industry: Insurance

The responses are broken down into a series of 'functional words' and 'qualifiers' as described above under section 6.4.2. The coder knows, by training, that the functional word is 'Manager' and the primary qualifying is 'Branch'. Abbreviated versions of these key words (three characters from each) are typed into the microcomputer, which generates the following 'tree' structure:

Manager, Branch insurance office:
- selling
- supervising staff, administration
- no other specific information

In this instance, the coder has been instructed to select 'Manager, supervising staff, administration' as the relevant index entry for the this particular response. The computer returns the code and automatically records the information. In other instances the computer may not return a code, but directs the coder to seek assistance.

The ABS system is probably still the best of the existing examples of the integration of coding instructions with a structured coding index and the resulting direct relationship between the index and the classification. The classification dictates the amount of index material which must be available if the coder is to classify occupational information correctly. The index cannot be extended simply by 'adding-in' new job titles, but requires that job descriptions be researched before extending the tree structure further. The coder must undertake a systematic review of all the occupational information available before entering the index, which then provides a guided route to the relevant occupational code. Through this procedure very high levels of consistency in coding are obtained.

This approach requires trained coders to operate the system. Also, the system does not, at first sight, appear 'natural' to the coder who may be more used to a process of matching job titles and index entries in a large alphabetically sorted index. This means that the system does not allow the management of the census or survey or the coding staff an easy way to compromise on the quality of coding to achieve lower coding costs through 'short cutting' the search procedures for response/code matches. From some points of view this may be seen as a disadvantage of the system.

6.6 The Cote d'lvoire Living Standards Survey

A continuing sample survey of 1,600 households per annum was undertaken by the Cote d'lvoire Department of Statistics in cooperation with the World Bank. A brief description of the use of computers in this survey is included, even though the collection of occupational information does not occupy a major role in the data collection process, because of the innovative use of microcomputers in the coding and editing of data by fieldwork teams.

The fieldwork operation for the survey consists of five teams stationed in regional offices of the Department of Statistics. Each team contains one 'data entry operator'. Interviewing of respondents takes place at the start and end of a two week period, allowing the team time to collect information from households, key the information into a microcomputer, verify and edit the data and to return to the household with any queries raised by computer consistency checks from the first interview. Data entry is done via a keyboard with the computer program controlling data entry exhibiting entries on the screen to resemble the original questionnaires. (Occupational information was coded separately from the computerized system.) Computerization of data entry and editing in the field led to major gains in terms of the time that elapsed from the start of the survey to the publication of first results. Problems which the survey teams reported were relatively minor in nature and refer to the hardware in particular. There was a need to maintain the computers in an air-conditioned environment, to prevent dust damage and to provide voltage stabilizing equipment due to mains voltage fluctuation.

6.7 Concluding remarks on computer assistance

Although computer systems have the potential to yield dramatic improvements in the quality of data, to control the environment within which data are collected 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. The costs must include a realistic estimate of the rate of depreciation of hardware, possibly over a period as short as one or two years. The costs must also take account of the reliance on specialist programming and systems analysis skills for the development of the required software and the cost of training users to work in the disciplined environment of a computerized data collection and coding system. One approach to the development of computerized data collection and coding which would minimize the risks of development (though not necessarily the cost) would be to acquire the rights to operate a system already tried and tested. This has the advantage that the potential application can be gauged against an actual application.

7 The problems of different languages

In the discussion above no reference has been made to the problems encountered in countries where there is more than one language which the population use in their daily life. The only adequate way of dealing with this situation, from the point of view of data quality, is to create separate data collection and processing instruments for each language, i.e. have interviewers speaking the different languages as well as having separate questionnaires, instructions and coding indexes for each language. (With adequate coding indexes there should, in principle, be no need to translate group descriptions in the occupational dictionary from the original language. That is, at least not for the purpose of producing occupational statistics.) However, in most countries this is not a possible solution because of the costs involved. The choice of second best solutions must be based on the circumstances of the country, taking into account such factors as how different the various languages are; to what extent the occupational vocabulary is common, for example for jobs in the modern sector; the extent of knowledge of a common second language, such as Swahili and English in some East-African countries, French in some West-African countries and English in some Asian countries.

Using interviewers who do not speak the same first language as the respondents, or using questionnaires written in a language which they do not understand, greatly increases the risk that the respondent will not properly understand the question posed and therefore that he/she will give an answer which is irrelevant or misleading, or the wrong code (alternative) for the response will be chosen. It also increases the risk that the interviewer will not properly understand the answer given by the respondent, even if the answer is correct. In addition the possibilities for successfully probing for better responses are greatly reduced in such situations. An increased risk that either the question or the response, or both, is not properly understood will of course increase the risk that what the interviewer writes down for later coding by him/herself or an office coder will be misleading as a description of the respondent's job, and thus increase the risk that an incorrect code will be used.

The interviewer and the respondent may be able to communicate without misunderstanding, but it may not have been possible to develop a coding index in the language used by the respondent. In this case one possibility is that the interviewer translates the response into the language of the coding index. This he/she will automatically have to do, at least implicitly, if he/she is also coding. An alternative is that the interviewer writes down the respondent's answer and that someone else, for example the coder, translates this into the language of the coding index. The problem with these solutions 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 occupational terms are used in the context. Very few persons will normally be able to satisfy this requirement over the whole range of work situations covered by a population census or labour force survey.

When presenting its occupational statistics from a population census or survey the statistical organization must carefully evaluate whether and to what extent the results for certain geographic areas or for certain population groups may have been influenced by the problems of language, in terms of general quality of the results or in terms of specific biases. This is, of course, part of its general responsibility for providing adequate description of data quality, including explanations for possible shortcomings in the data for particular groups.

8 Mapping a national occupational classification to ISCO

The purpose of this chapter is to outline a strategy for countries that wish to map a national occupational classification (NOC) into the revised International Standard Classification of Occupations (ISC0-88). The focus is on mapping the groups of a NOC into corresponding groups in ISC0-88. The discussion is based on the assumption that the NOC is based on principles roughly consistent with those of ISC0-88, in particular, that the most detailed groups are defined in terms of type of work performed.

We begin with a brief outline of why it may be useful and important to establish links between NOCs and ISC0-88. The next section outlines why and how mapping should always be carried out at the most detailed levels possible of the classifications. Then a short discussion follows of the procedure for establishing links at more aggregate levels and of the use of double coding. The last section outlines the role of the ILO in the establishment of links between NOCs and ISCO-88.

8.1 Motivation

National occupational classifications should be mapped into ISCO-88 mainly because national users of occupational information want to:

(a) make comparisons between national circumstances and circumstances of other countries;

(b) communicate occupational information with persons or institutions in other countries.

An additional consideration, but less important, is that users in other countries and in international organizations also want to make comparisons between countries and communicate occupational information internationally.

If only two countries are involved, the need for international comparable occupational information could be satisfied most effectively by directly linking their national classifications. However, as soon as more than two countries are involved, pair-wise linking becomes very cumbersome. Even if most comparisons are expected to be pair-wise, it may be more efficient to use the indirect route of linking to a common reference classification in order to avoid having to establish many pair-wise links. The obvious candidate for the role of a common reference classification is ISCO and, increasingly, its last version - ISCO-88.

It is important to remember when discussing international use of occupational information that both statistical and client- oriented use should be considered. As ISCO is intended to facilitate international occupational communication for both client-oriented and statistical users, ISCO-88 tries to provide a basis for the different uses at the national level while taking into account the special considerations which must follow from its international nature.

Internationally comparable statistics on occupational groups are used mainly to:

(a) compare the distribution of the employed population or some other variable (such as wages, hours of work, work accidents, income, consumption, reading habits) over occupational groups in two or more countries;

(b) compare data on broadly or narrowly defined individual sets of occupations in two or more countries, for example, to compare the average wages of computer programmers in country "A" with those in country ~B", or the number of “industrial designers" in the two

countries;

(c) merge data from different countries referring to comparable groups - in order, for example, to obtain enough observations to study the incidence of particular work related accidents or diseases among workers believed to have similar exposure to harmful working conditions or substances.

Depending on the purpose of the study, “occupation" may be regarded as the main variable, or it may serve as a background variable in a statistical analysis. Used as a background variable, it sometimes serves as a proxy for other variables, such as “Socio-economic groups" or “working conditions", or it is used in the construction of other variables. Experience shows that at the international level many users of occupational statistics need data mainly at the higher level of aggregation - usually for type (a) descriptions. Among the exceptions are international studies of wages and earnings, work hazards and injuries and other conditions of work. Such studies often require that detailed occupational groups are defined, sometimes in cross-classification with the “industry” and/or “status in empIoyment" variables.

The main client-oriented applications for a standard international occupational classification are in the international recruitment of workers and in the administration of short- or long-term migration of workers between countries. An internationally developed and agreed upon set of descriptions for detailed occupational categories which can serve as a common “language" for the countries and parties involved in such programmes may greatly increase the effectiveness of the communication necessary for their implementation.

While the statistical use of type (a) above requires that the occupational classifications cover all jobs, the focus in other types of use - statistical or client-oriented - is on specific occupations or groups of occupations. Although the sum total of all users interests in these types of use could conceivably also cover all occupations, in practice only some occupations are involved. The problem is to know which they are. Therefore the links have to be established for the whole range of occupations covered by the two classifications.

Because occupational information is needed at all levels of aggregation in international, as well as in national, applications, it follows that links between a NOC and ISCO-88 should also be established as far as possible at all levels of aggregation. In order to achieve this it is necessary, but not sufficient, to establish links at the most detailed level in the classifications. However, the links established at the detailed level will ease the establishment of links at more aggregate levels.

8.2 Mapping at the most detailed level

Mapping one classification into another is equivalent to determining for each group in the first classification the most appropriate group in the other. This is in principle very similar to coding an occupational response on a questionnaire - an advantage being that in the case of mapping, one normally should have access not only to a title and very brief task information when assigning a code, but to a whole description of tasks and duties of jobs for each occupational group included in the two classifications, i.e. the NOC and lSCO-~R

The first step when establishing links should always be to give to the most detailed groups of the NOC the code of the most detailed appropriate group in ISCO-88. Assuming that the most detailed level in the NOC is more detailed that the unit group level in ISC0-88, we will have the following cases:

(a) The NOC group belongs unambiguously to one of the ISCO-88 unit groups. This is of course the simplest situation and, if the NOC was developed on the basis of ISCO-68 or ISC0-88. this is likely to be the most usual case;

(b) The range of tasks and duties of the jobs belonging to the NOC group is partly outside those described for the most relevant ISCO-88 unit group, but falls within the same ISCO-88 minor group. In this case the group should be coded according to the numerical dominance priority rule outlined below, or, if this is not applicable, to the appropriate minor group;

(c) The range of tasks and duties of the jobs belonging to the NOC group is partly outside those described for the most relevant ISCO-88 unit group and they also fall partly outside the corresponding ISC0-88 minor group. In this case, the group should be coded to the ISCO-88 unit group determined by following the priority rules outlined below, invoked in the same order as they are described. If no unit group can be determined, then the same exercise should be carried out to determine the most appropriate ISC0-88 minor group, sub-major group or, as a last resort, major group.

The numerical dominance priority rule would say that in the case of a detailed NOC group not fitting into any ISC0-88 unit exactly, the group should be coded to the ISCO-88 unit group to which a large majority of the jobs, (around 80 per cent) in the national group belongs. If there is no such ISC0-88 unit group, then one should try to use the skill level priority rule.

The skill level priority rule would say that the national group should be coded to the ISCO-88 unit group which includes those of the tasks and duties of the national group which correspond to the highest ISC0-88 skill level. If no difference in skill level is involved in the different tasks and duties. then one should try to use the production priority rule.

The production priority rule would say that priority, when deciding to which ISCO-88 unit group to code, should be given to those tasks and duties which are directly related to the production of goods or services rather than to associated tasks and duties related to the sale and marketing of the same goods, their transportation or the management of the production process (unless any of these tasks and duties predominates among the workers in the NOC group). For example, when the tasks are baking bread and pastries and also selling them, the priority should be given to baking, not to selling; if the tasks are operating a particular type of machinery and also instructing new workers in how to operate the machine, then priority should be given to the machine operation; if the tasks are driving a taxi and also keeping the accounts, then priority should be given to driving.

The result from this coding exercise would be a list of detailed NOC occupational groups mostly given ISCO-88 unit group codes or the codes of even more detailed ISCO groups in areas where such groups may have been defined. This means that, when needed, the NOC groups can subsequently be aggregated to most aggregate groups which have been defined in terms of ISC0-88 unit groups.

By coding a random sample of responses from a population census or survey to both NOC and ISC0-88, it is possible to get a picture of how jobs belonging to a particular group in a NOC are distributed over ISC0-88 groups. The results can then be used for statistical mapping, i.e. to distribute the number of jobs or persons of a NOC group among the relevant ISC0-88 groups. However, the costs of double coding may be a significant problem, unless the coding is done by recording the entry (line) number of a coding index where each entry has been given both codes. In this case we can talk about simultaneous coding to two (or

more) classifications rather than about double coding.

It should be noted that even though the results from double coding usually will be the best source of information for determining whether the numerical dominance priority rule can be applied, other sources and more unstructured general knowledge about the labour market and the population being coded may supplement or replace the results of double coding.

8.3 Mapping at aggregate levels

Unfortunately statistical data very often are not available for groups defined at the most detailed level in the NOC, mainly because the census or survey returns have been coded and/or tabulated at a higher level in the NOC. This makes it necessary also to establish links directly between the aggregate NOC groups and the most detailed relevant aggregate ISCO-88 groups. The first step in this process should be to look at the structure of the aggregate NOC groups for which data are available in terms of their component ISCO-88 unit groups, i.e. at the results of the exercise described in the previous sections. Using the same priority rules as those outlined above, one should determine how one or the sum of several NOC aggregate groups can be used as a reasonably close approximation of an ISCO-88 unit group, minor group or sub-major group. In terms of closeness of approximation, this procedure evidently will give results which are much less satisfactory than those resulting from aggregating data using detailed national groups. This is one reason why it is recommended to always code to the most detailed level of the national classification, given the information in each census or survey response.

8.4 The role of the ILO

Mapping a NOC into ISC0-88 requires a good understanding of the national labour market and occupational structure, of the NOC itself and its principles, and of ISCO-88. Those who are responsible for the NOC and who have experience with its use are therefore best suited to establish the links between NOC and ISCO-88. However, because the exercise involves both classifications, it would be an advantage if those responsible for ISC0-88 could be given the opportunity to comment upon a first draft of links between the two classifications, based on their knowledge of ISCO-88 and their experience with the difficulties of other countries and the way these have been resolved. Contacts should therefore be taken with the ILO Bureau of Statistics before finalizing the links between the NOC and ISCO-88.

For NOCs which were originally developed on the basis of ISCO-68 - or to a lesser extent, ISCO-58 - the links established between ISCO-68 and ISCO-88 at the detailed level through the index in the ISCO-88 publication may prove useful. The index can also be made available to countries in machine-readable form.

9 Summary:

Twelve golden rules for capturing and processing occupational information

1 Interview the job holder.

2 Use his/her first language.

3 Define clearly which job(s) you are asking about.

4 Ask a question (or questions) designed to obtain information both about the job

1. title and about main tasks and duties of the job.

5 Train the interviewers to understand and write down what the coders need to know.

6 Train the coders to understand what to select from the response and how to match this with the coding index.

7 Design the coding index to reflect both the type of responses which will be given and the coding instructions.

8 Use the information in the responses in this order of priority: title, tasks, industry/employer (only for the cases specified), qualification (only by supervisors).

9 Code to the most detailed level of the classification supported by the information in the response.

10 Give interviewers and coders feed-back on performance.

11 Use systematically the queries generated during the coding process to update the coding index and the classification.

12 Mapping from the NOC to ISC0-88 should always be made between the most detailed groups possible.

Suggestions for further reading

Australian Bureau of Statistics (1990): *Australian Standard Classification of Occupations {ASCO) - Expert Coding System: Unit Group Level, Version 5.0 on Floppy Disk.* Canberra, Catalogue no. 1224.0

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

Australian Bureau of Statistics (1992): *Australian Standard Classification of Occupations {ASCO) - Manual Coding System: Unit Group Level.* Canberra, Catalogue no.1225.0

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

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

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

Embury, B.L. (1995): *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 2/1995. 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, l., ea.: *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, London, New York.

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. t1973): "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.

United Nations' Educational, Scientific and Cultural Organization (1976): *International Standard Classification of Education - ISCED.* (COM/ST/ISCED), UNESCO, Paris.

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

United States' Bureau of the Census (1982): *Classified Index of Industries and Occupations: 1980 Census of Population. US* Department of Commerce, Washington DC