Table 5

Table 5 presents population by language, sex and urban/rural residence for each census between 1985 and 2003.

Description of variables: Statistics presented in this table are from population censuses. Statistics shown pertain to three major types of language data. These are: (1) mother tongue, usually defined as the language spoken in the individual's home in his or her early childhood, (2) usual language, defined as the language currently spoken, or most often spoken, by the individual in his or her present home, and (3) the individual's ability to speak one or more designated languages. Information on all languages spoken by each person is the basis of the classification for one or two distributions.

Language is one of five separate characteristics (country of birth, country of citizenship, ethnic group, religion and language) which can be used to explore some facet of the ethnic composition of populations. Language, and particularly mother tongue, is probably a more sensitive index for this purpose than either country of birth or country of citizenship because linguistic differences tend to persist until complete cultural assimilation has taken place. Common ancestral customs may be reflected in the mother tongue of individuals long after these persons have changed their citizenship. Thus, important ethnic groups, not only among foreign-born alone but also among native-born or second generation population groups, may be distinguished by language differentials.

Data on the language currently spoken are somewhat less appropriate for the identification of ethnic groups in the population, but they are adaptable to other uses, such as the investigation of the linguistic assimilation of immigrant groups. For this purpose, however, there is need for corresponding data on country of birth or citizenship. Their value in this connection is enhanced by the availability of data from a series of censuses.

Statistics based on ability to speak a specified language or languages are perhaps the least useful for identifying ethnic groups, but they do have utility in connection with problems of educating and communicating with linguistic minorities, and they serve as a means of pointing out the heterogeneity of languages within a country and of measuring their individual strength. Such data are especially important in countries where more than one official language is recognized.

The United Nations Principles and Recommendations for Population and Housing Censuses Rev. 1 (Sales No. 98.XVII.8) recognizes the different purposes served by each of the above kinds of data and point out that each country should collect the type of information most appropriate to its need. The particular kind of data shown in each tabulation is, therefore, identified in footnotes, along with a concise summary of the definition used by the country, when known.

Regardless of the type of data shown, the categories in each distribution are arranged in English alphabetic order. "Other" and "Unknown" are at the end.

The urban/rural classification used is that provided by each country or area. National definitions are given in the regular issue of the Demographic Yearbook in the notes for table 6.

Reliability and limitations of data: Errors in national census data can arise at any stage of the collection, processing or presentation, and such errors may limit the quality and international comparability of census statistics presented in the Demographic Yearbook. Two major types of errors in census data are often distinguished: first, coverage errors, which lead to the over-enumeration or under-enumeration of the population in the census, and second, content errors, which affect the accuracy of the recorded information for the covered population. Because coverage errors may occur more frequently among some population subgroups than others, coverage errors may affect not only the absolute number of persons in any given category but also their relative distribution. Levels and patterns of coverage and content errors differ widely among countries and even, at times, from census to census for a specific country. Further limiting the international comparability of census statistics are variations among countries in the concepts, definitions and classifications used in their censuses.

The obvious limitation in data on language for international comparison purposes is that the statistics refer to three separate types of data and that these three are not strictly comparable. Data on mother tongue are designed to identify cultural or ethnic groups in the population by means of the language reported to have been spoken in early childhood— presumably before immigration, if that is a factor. They do not indicate linguistic ability, but rather a cultural group as defined by language.

Information on usual language, on the other hand, reflects linguistic ability at the time of the census and may mask the fact that persons using the same dominant language of the country in which they reside will usually include a diversity of ethnic groups.

Questions on ability to speak a specified language will not ordinarily produce statistics comparable with either of the other two types.

In addition to the basic lack of comparability between the three different types (mother tongue, usual language, and ability to speak one or more languages), attention must be called to the limitations of each type of data. The concept of mother tongue, the best of the three for analysing ethnic composition, produces more or less comparable data from country to country. The census question on which the statistics are based is essentially the same in each country, except for the occasional requirement that the language reported must still be understood by the person. Questions on language currently spoken, however, may take various forms, including a request for language usually spoken, language best spoken, language spoken fluently, language spoken with family, language spoken in addition to mother tongue, and so forth. The manner in which the question is asked and the criterion of age adopted contribute to a wide degree of variation in the answer to this type of question possible.

The age limits chosen for tabulation of data on usual language or ability to speak specified languages introduces problems in comparability of data. Where no age limit is set, young children are usually assigned the language reported by the parents, or, in the case of mother tongue, the language currently spoken in the home. Almost all of the distributions in this table are of this type, that is, they are for persons of all ages; exceptions are footnoted.

It may be noted that the treatment of multilingual persons may also introduce lack of comparability. In some censuses, speakers of more than one language are so tabulated, and all the various combinations and permutations of languages are set forth in the classification. In others, the major language of current speech alone is shown. These create primarily problems of tabulation and, so long as the population is counted and not the languages, the results may be compared to some degree. It should be noted, however, that occasionally, multilingual persons are counted separately for each language they speak, and the data, therefore, show the frequencies for languages, rather than the major language, or other combinations of languages, spoken by each person; hence, the sum of the frequencies in such cases is greater than the total population.

One of the most important limitations on the use of population classified by language is the lack of uniformity in the detail shown in any of the three types of classifications. The United Nations has recommended tabulating separately all languages of numerical importance in the country. The problem arises in defining a language as distinct from a dialect and in determining what is numerically significant. For some countries, great detail is available while, for others, a large "other" or residual category is included.

Finally, it should be emphasized that data in this table do not lend themselves to determining how many persons in the world speak a certain language. Not only do the definition and the amount of detail differ from country to country, but not all countries are included in the table. For many countries or areas, language is not a subject of investigation.

The comparability of data by urban/rural residence is affected by the national definitions of urban and rural used in tabulating data. The definitions vary considerably from one country or area to another.

Earlier data: Population by language has been shown previously in issues of the Demographic Yearbook featuring population census statistics as the special topic. Most recently these data were published in the 1993 Demographic Yearbook.