Working Paper No. 3
8 November 2004

ENGLISH ONLY

STATISTICAL COMMISSION and UN ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Seminar on New Methods for Population Censuses
Organized in cooperation with UNFPA
(Geneva, 22 November 2004)

Session 1 – Invited paper

THE 2008 ISRAEL INTEGRATED CENSUS OF POPULATION AND HOUSING

Submitted by Central Bureau of Statistics, Israel*

The following note describes the principal components of the 2008 Israel Integrated Census of Population and Housing, to be conducted by the Israel Central Bureau of Statistics (ICBS). Section A presents the plans for the integrated census; Section B addresses the issues specified in the agenda for this session. (A different version of Part A appears as Document ESA/STAT/AC.98/9 in the UN Expert Group Meeting to Review Critical Issues Relevant to the Planning of the 2010 Round of Population and Housing Censuses, 15-17 September 2004, United Nations, New York.)

Section A

I. BACKGROUND

1. In addition to its basic function of providing population counts, the Israeli census has four principal functions, which are similar to those characterizing censuses everywhere: (1) the demographic information on the population is used by the Central Bureau of Statistics to compute the weights applied to estimates based on our sample surveys of the population in order that they reflect the characteristics of the population; (2) enumeration of the population in localities provides the Ministry of Interior and other ministries with the information they require for budget allocations; (3) data from the 20% sample is the sole source of detailed socio-economic information for the entire population by geographical division and sub-groups; (4) making data available to users for a wide range of purposes: government, commercial, academic, educational, research.

* Paper prepared by Charles S. Kamen

2. Israel's current population is 6.8 million. Israel has conducted five population censuses, in 1948, 1961, 1972, 1983 and 1995. The first census, in 1948, had the primary function of establishing the Population Registry (PR). There is no legal requirement in Israel for conducting a population census at predetermined intervals, nor one requiring that a census be conducted at all. That is the reason for the irregularity of the census dates. A specific government decision is required in order to conduct a census, and in practice it is the ICBS that initiates the process resulting in this decision. The arrangement leads to uncertainty in census planning and hinders the establishment of an ongoing census unit with a multi-year work program. For example, our forthcoming census was originally scheduled for November, 2006, but the government decided to postpone it for two years.

3. In 2008, we plan to move from a "conventional" to an "integrated" census. In a conventional census the goal is to physically enumerate every person, either directly or by proxy through a member of their household. The questionnaire for the 100 percent enumeration usually contains a small number of basic demographic questions. A longer questionnaire is usually administered to a sample of those completing the short form. The main reason for restricting the long form to a sample is to reduce costs. All Israeli censuses except for the first used a short questionnaire for everyone and a long questionnaire for a 20 percent sample of households. The 2008 integrated census will enumerate only a 20 percent sample of households; there will be no complete enumeration of the population.

4. Conventional censuses have two major shortcomings: They are expensive, and are very vulnerable to lack of cooperation by the public. Not only are they expensive: their cost can be expected continually to increase, for three reasons. First, the population increases, and since a substantial portion of the cost of a census goes for enumeration, the increase in the size of the population raises total cost of enumeration even if the cost per capita remains the same. Second, salaries rise, and since a major component of census costs are salaries, the total cost of the census will increase even if the population doesn't increase. Third, censuses will become increasingly dependent on computer technology. The technology used in one census will most likely not be used in the following census a decade or more later: it may no longer be available, or even if available may have been superceded by technologies which provide more for the money - though they also cost more.

5. Public cooperation with the census is also likely to decline, for at least four reasons which are interconnected: resistance to government intrusion; increasing awareness of, and concern about, privacy; growing individualism which makes people less willing to participate in communal endeavors such as a census; and the policy of privatizing public services which undermines community.

6. The obvious alternative to a conventional census is one that is "register-based," and utilizes official lists of persons created for administrative purposes. There are many examples of such lists: births; driver's licenses; pupils in school; tax filers; dwellings; social security contributors; etc. The "best" official list for census purposes is a Population Registry that contains information on everyone in the population; many countries, however, lack such a registry. Israel has a registry and each person in the PR has a unique PR ID number. In theory, Israel's Population Registry could provide the same information as the short census form: age, sex, address, place of birth, date of immigration, race/ethnicity, marital status, religion, kinship relations.

7. Why, then, do we need a census at all? Israel conducts a de jure census. It counts people at their usual place of residence. At the national level this is defined, for persons listed in the

Population Registry, as those present in Israel or absent for less than one year; and for those not listed in the Population Registry, as persons who have lived in Israel for one year or longer. Israel's Population Registry cannot now substitute for a census, for four reasons: First, the Population Registry is not coterminous with the list of persons defined as comprising the de jure census population. It contains persons who are not part of the de jure population, in particular emigrants who no longer live in Israel. Moreover, it does not include persons lacking PR ID numbers who are resident in Israel continuously for a year or longer, legally or illegally. Second, the geographical information in the PR, and in particular the addresses, is of poor quality: approximately one fourth of the persons in the PR are listed at addresses other than where they actually live. Third, the PR doesn't include the socio-economic information obtained on the census long form. Fourth, the PR does not include information on housing.

8. The goal of the integrated census is to combine the benefits of a conventional census with those of a register-based census, while reducing the shortcomings of each. The integrated census is designed to provide population counts and estimates of population characteristics by combining information from the Population Registry with that obtained from a sample survey in the field. One of the main reasons that we have to conduct an integrated census is because Israel has no register of dwellings; thus, we are unable to allocate households to dwellings on the basis of administrative information. The sample survey enables us to provide information on households.

9. The integrated census will provide estimates for localities and, within localities, for statistical areas (equivalent to "census tracts"). We do not, at this point, envisage that we will be able to provide estimates of socioeconomic characteristics from the sample enumeration for geographical areas which are smaller than statistical areas or whose borders are not congruent with those of statistical areas or localities.

II. The conception underlying the integrated census

10. The conception behind Israel's 2008 integrated census is simple. We take the information in the Population Registry as a first estimate of the size of the population and its geographical distribution. We undertake two parallel operations. (1) We update the address information in the Population Registry on the basis of address information from other administrative data files. (2) We select a cluster sample of contiguous addresses containing either single or multiple dwellings for enumeration of the households living in them. The information obtained from this sample enables us to evaluate the coverage of the PR and to collect the demographic and socio-economic data on persons and households that is obtained in the sample questionnaire of a conventional census. The estimation procedure extends the dual system model for estimating undercoverage, in order to provide an estimate of overcoverage.

11. The basic census procedure is as follows. Using the existing administrative-statistical division of the country into localities, and statistical areas within localities having 10,000 inhabitants or more, we divide them into Enumeration Areas (EA), groups of 50 geographically contiguous households whose boundaries do not cross boundaries of localities or statistical areas. The division into EAs is based on address information in the Population Registry. The PR is a list of individuals, each of whom has a unique ID number. It is not a list of dwellings, nor a list of households. But the PR record for each individual often includes the ID number of his or her children and spouse. It is therefore possible to create "administrative households" within addresses ("administrative", because the individuals listed don't necessarily live at that address, or in the same apartment) and use them to define the EAs. We sample approximately twenty percent of the EAs within each statistical area (or within the locality if it has fewer than 10,000

inhabitants), taking into consideration various population characteristics in order to make the sampled EAs as representative as possible of the statistical area (or of the locality). We enumerate all the households found in the dwellings at the addresses located in each EA. For each "actual" (as opposed to "administrative") household we find in the sampled EA, we fill out a long questionnaire. At the end of the enumeration, we have for each EA two lists of persons: those who were listed in the updated Population Registry at the addresses included in the EA, and those whom we actually found in the field enumeration. We generate these two lists independently: by that I mean that the enumerator is not provided with a list of persons or households who, according to the PR, live at the addresses he has to cover, and hence the list of persons which he or she generates as a result of the enumeration is independent of the list of persons in the PR.

12. I will distinguish in what follows between "registry **overcoverage**," which refers to persons listed in the PR who do not **actually** live in the sampled EA, and "registry **excess**," which refers to persons who were not **found** in the sampled EA. Registry **excess** is observed; registry **overcoverage** is inferred. Each person at a given address in the sampled EA falls into one of three categories: (1) listed at that address in the Population Registry, and actually enumerated at that address in the 20 percent sample survey; (2) listed at that address in the Population Registry, but not enumerated at that address in the sample survey - "registry excess", compared to the field; (3) enumerated at that address in the sample survey, but not listed there in the Population Registry. These are "field excess," compared to the PR.

13. I noted above that we update the address information in the Population Registry on the basis of other administrative data files. We do so to reduce the number of persons in the category of "registry excess." This is important because we will, in the next stage, attempt to locate them. If we can reduce their number by updating their addresses (thereby probably removing them from the PR list for a particular EA and adding them to the PR list for another EA), we won't have to spend time and money looking for them unnecessarily. In addition, by reducing the amount of excess we will improve our estimates.

14. Since the basis for the integrated census is the updated PR, we have to evaluate its errors - these errors are the PR overcoverage and undercoverage. To carry out this evaluation, we must locate the people who were not enumerated at their PR address in order to find out where they actually live. Although the sampling unit for the integrated census is the EA, the census itself is intended to provide estimates not for EAs, but for localities and statistical areas. Thus, in our search for the persons categorized as "excess", it is sufficient to place them in the appropriate statistical area.

15. The result of our comparisons among the list of persons we enumerated in the EA, those whom we located after the enumeration and those listed in the PR for that EA allows us to classify the persons on the PR list into five outcome categories: (1) persons listed in the PR at the address in the statistical area at which they lived; (2) persons who live at an address which is not in the same statistical area as the one in which they are listed in the PR for the sampled EA, but is in the same locality; (3) persons whose "actual address" is not in the same locality as the PR address, but is in Israel; (4) persons whose actual address is not in Israel; (5) persons whose actual address is unknown to us (these persons will be allocated among the other four categories).

16. By this point we have obtained information on the location of each of the persons appearing on the updated PR list for the EA. We use the information about the results of our search for the persons appearing in the PR lists for the sampled EA's in each statistical area in order to infer what the results of the search would be were it carried out for all the persons listed

in the PR for a given statistical area. We compute the probability that a person listed in the updated PR in the sampled EA's of the statistical area falls into one of the five "outcome categories" listed in Par. 15, on the basis of the information obtained from the field survey and the search for "registry excess."

17. In a field test conducted in Bet Shemesh (a town near Jerusalem) in 2002, these outcome probabilities were computed separately for each of four age groups within each statistical area: 0-20; 20-30; 30-40; 40 and older. The result of this procedure is as follows: Each person listed in the updated Population Registry is assigned a probability of being in one of the four (after allocation of those whose address is unknown) outcome categories according to his age and statistical area of residence. We use these probabilities to estimate the registry overcoverage. The undercoverage is estimated on the basis of the field enumeration. Combining the overcoverage and undercoverage estimates, and taking into consideration the sampling fraction, we compute the census weight for the PR record. The population estimate for any population group is the sum of the PR weights assigned to its members.

III. STATISTICAL METHODOLOGY OF THE INTEGRATED CENSUS

18. The statistical methodology of the integrated census has two main components: estimating coverage of the population, and estimating population characteristics. Both the coverage estimates and the estimates of population characteristics use information from the updated PR file and from the field enumeration survey. The coverage estimates are based on the methodology of Dual System Estimation, and represent an extension of the classic model for estimating undercount in census data; the extension accommodates "false captures", or overcount, in the administrative list. [for a more detailed presentation of the model, and of its mathematical basis, cf. Glickman, H., Nirel, R. and Ben Hur, D. (2003), "False captures in capture-recapture experiments with application to census adjustment." Bulletin of the International Statistical Institute, 54th Session, Contributed Papers, Vol. LX, pp. 413-414; Nirel, R., Glickman, H. and Ben Hur, D. (2004), "A strategy for a system of coverage samples for an integrated census." Proceedings of Statistics Canada Methodological XX Annual Symposium (forthcoming)]. The lists generated in the classic "capture"/ "recapture" stages are represented in our model by the two independent lists we create - one based on the updated PR, and the other on the field enumeration.

19. Our procedure produces population estimates for each statistical area. We define the PR coverage errors with respect to a statistical area. The PR **undercoverage** for a given statistical area is composed of all persons living in that area but listed elsewhere in the PR. The PR **overcoverage** for a given statistical area is composed of all persons listed in the PR in that area but not living there. The field enumeration described in Par. 11 and the procedure for locating "registry excess" described in Par. 13-16 enable us to generate for each statistical area entries for three of the four cells in the following fourfold table:

	Enumerated in the SA	**Not enumerated in the SA**
In the PR for the SA	(1) PR identical to Field	(2) PR excess
Not in the PR for the SA	(3) Field excess	(4) Possible undercoverage

Persons listed in the PR for a given statistical area but living in another statistical area or abroad are excluded from the dual system table for a given statistical area. These persons comprise a "fifth cell" representing the PR overcount (We assume the field enumeration data have no false captures). Using the entries in the fourfold table, and assuming independence between the two

lists, we estimate the probability that a member of the population will appear in the PR list. This provides the registry undercoverage parameter. Using the "fifth cell" we are able to estimate the proportion of overcoverage in the PR, thereby obtaining the registry overcoverage parameter.

The entries in cells (1), (2) and (3) are obtained from the field operations. The entry is cell (4) is the census undercoverage, which we estimate. Therefore, the sum of the four cells, which equals the total number of persons who live in the statistical area, is also an estimate. It will be necessary to evaluate the quality of our estimate of the census undercoverage.

20. Using the numbers in the fourfold table, and assuming independence between the two lists, we estimate the probability that a member of the population will appear in the PR list (the estimate of registry undercoverage). As noted earlier, we know there is overcoverage in the updated PR file, of persons listed in one statistical area who in fact live in another statistical area or abroad. Using the sample of EAs from the PR we are able to obtain an estimate of the number of persons who are registry overcoverage.

21. The extended dual system estimates are based on the following assumptions: (1) independence between the PR list and the field enumeration list; (2) homogeneous capture probabilities - all persons in the population have the same probability of being included in the PR, as well as the same probability of being enumerated in the field; (3) the distribution of registry overcoverage across EA's is proportional to the distribution of the "true" population. The assumption of independence is met because the field enumerators have no knowledge of the contents of the PR. The two other assumptions are met by dividing the population into homogeneous groups with respect to the likelihood that they are subject to errors of overcoverage and undercoverage, basing this division on variables related to coverage error.

22. The estimation procedure proceeds by the following steps: (1) establishing estimation groups homogeneous with respect to coverage errors; (2) computation of the overcoverage and undercoverage parameters for each group; (3) computation of a "census weight" for each person in the final census file according to the estimation group to which he belongs. The "census weight" is the coefficient by which the PR record is multiplied in order to represent the number of persons it represents in the population. At the end of this procedure we are able to compute population estimates for subgroups with various combinations of characteristics, and to create a final census file with each record assigned a census weight.

IV. GENERATING THE DATA REQUIRED BY THE INTEGRATED CENSUS

23. As described above, the integrated census involves two main stages. In the first stage, two lists are created for each EA - one of persons listed in the updated Population Registry, and the second of persons enumerated in the field. In the second stage, we try to locate those on the PR list ("registry excess") who were not enumerated in the field. Our efforts to update the PR focus on two components: address, and presence in the country (which determines whether a person should be included in the census population). Addresses are corrected by comparing the content of the PR file with the content of three other administrative data files - driver's licenses; electric company accounts; pupils in elementary and secondary schools - using the person's unique PR ID number to link individuals across data files. Presence in the country is corrected by using information from the border control files. In addition, PR update files are obtained in order to identify births, deaths and marital status changes which have occurred since we received the previous PR file; they also include new immigrants.

24. Field enumeration involves three basic stages: (1) a pre-enumeration canvass of the EA, during which the enumerator lists dwelling units; (2) enumeration, during which the enumerator returns to the listed dwelling units and interviews the residents; (3) a "clean-up" stage, in which the enumerator makes a final effort to enumerate dwellings that were closed and to convert non-response. The field enumeration has two goals: to obtain socio-economic information on the population, and to obtain the information needed to evaluate the address information in the updated Population Registry so it can be used as the basis for census estimates.

25. We obtain information on the location of persons in the registry excess group in a number of stages. In the first stage, the enumerators return to their EA's with lists of persons whose PR address is in the EA but who were not enumerated, and try to locate them - either by actually finding them in the field, or by obtaining from others information on their location. In the second stage, the names of persons not located by the enumerators are transferred to a CATI system, and information on them is sought by phone. Additional potential sources of information on the location of registry excess persons are updated versions of the administrative data files used to correct the Population Registry, as well as other administrative data files that were not used for correcting the PR.

26. At the end of this search process we are, in principle, able to assign each person listed in the PR a probability that his PR address reflects where he actually lives, and if it does not, a probability that his address can be characterized by one of the four other alternatives listed above (Par. 15). This is the basis for creating the final 100 percent census demographic file. The 20 percent sample survey of the EAs is the basis for the final census file containing demographic and socio-economic information on persons and households.

V. 2008 CENSUS TECHNOLOGY

27. The 2008 Israeli census is computer-based in almost all aspects of its operations. Computer technology is what has made the integrated census feasible. The main computer-based components of the census include mapping applications involving preparing maps for enumerators, geographical anchoring of addresses and creating the network of EA's; harmonizing administrative files obtained from outside the ICBS; creating the updated Population Registry file by merging administrative files, including the development of algorithms for record linkage among those files; computer-based field work, including enumeration by means of laptop-based CAPI using Blaise software with computerized questionnaires in Hebrew, Russian and Arabic; devising an innovative solution that enables us to display fonts for these three languages within Blaise; questionnaire data transmitted by enumerators via the internet from their home telephones; field work administration based on data transmitted by enumerators via the internet from their home telephones; field staff management, hiring, allocation and payment integrated with census geography and enumerator production; training of field staff, including computer-based training materials and procedures; dissemination of results, including the development of applications for web-based table generators, summary tabulation generators and tables available on the internet.

VI. ENUMERATING SPECIAL POPULATIONS

28. The dual-list enumeration procedure described above is appropriate for approximately 70% of the population that lives in urban localities having an organized network of named streets and numbered buildings that can be mapped onto the addresses in the Population Registry. The others either live in localities without an organized system of addresses, live outside localities, live under arrangements which don't permit their enumeration by the standard dual list procedure,

or are not listed in the Population Registry. We are in the process of deciding for which of these populations it is necessary to develop special enumeration procedures: (1) large Arab localities having more than one statistical area and without an organized system of addresses; (2) small localities (both Arab and Jewish) having only one statistical area and no organized system of addresses; (3) persons in communal quarters (institutions); (4) residents of kibbutzim (collective settlements); (5) persons living outside the borders of localities; (6) concentrations of Bedouin living outside the borders of localities in southern Israel in the Beer Sheva area; (7) foreign workers; (8) the homeless.

VII. WHERE DO WE STAND NOW? 2004 FIELD TEST AND 2006 DRESS REHEARSAL

29. Field tests are conducted as part of census planning. These tests are particularly important in the case of the integrated census since both its conception and many of its procedures are new. Testing of the 2008 census procedures involves three main components: the behaviors required of the enumerators and other field staff in order to obtain census information; the process of creating the integrated administrative file used for the census; and the functioning of the computer programs and technology which underlie the work of the field staff.

30. The first field test, carried out in Bet Shemesh (a town of 50,000 inhabitants located 30 kilometers west of Jerusalem) in the spring of 2002, had four goals: (1) creating an improved address file (IAF) on the basis of the Population Registry and additional administrative files; (2) carrying out a field survey to obtain the information needed to correct the IAF; (3) on the basis of the field survey, creating a final weighted census data file; (4) computing population estimates using the final census data file. The results of the Bet Shemesh test showed the need to improve the procedures for locating the registry excess. That was set as a principal goal for the next field test in 2004.

31. The November, 2004, field test will be carried out in five localities: Giv'atayim (47,000 inhabitants, bordering Tel Aviv), Tira (an Arab town with 20,000 inhabitants near Tel Aviv), Yarhiv, Newe Yamin and Elishama (three small Jewish localities each having fewer than 1000 inhabitants). The goals of this field test are to evaluate: (1) procedures developed for locating the registry excess; (2) enumeration procedures developed for large Arab localities that don't have an organized system of addresses; (3) enumeration procedures in small localities that don't have an organized system of addresses; (4) on a small scale, procedures developed for managing field work; (5) on a small scale, the functioning of one local field office; (6) on a small scale, the effectiveness of the procedure for recruiting enumerators.

32. The third field test, in the fall of 2006, will be the major field test for the 2008 census. Our current plan is to include localities with a total of some 600,000 inhabitants, approximately 10 percent of the population. The test will be conducted in clusters of localities in two separate geographical regions, north and southeast of Tel Aviv. Strictly speaking, it will not be a "dress rehearsal," because it will include components which are being field-tested for the first time. The main goals of the 2006 field test are: (1) to test enumeration procedures for special populations; (2) to evaluate the functioning of the procedures for identifying and locating registry excess in localities varying in the likelihood of their residents being listed at their actual address; (3) to test the final versions of the enumeration procedures implemented in the 2004 field test and revised on the basis of the results of that test; (4) to test on a large scale the functioning of field offices, both vis-à-vis the field staff in each office and vis-à-vis census headquarters at the ICBS in Jerusalem; (5) to evaluate the procedures for arriving at census estimates; (6) to test preliminary versions of the internet-based data-dissemination tools.

VIII. MAKING THE RESULTS ACCESIBLE AND COMPREHENSIBLE TO THE PUBLIC

33. A variety of means exist for making the census results available to the public. These means have to meet two kinds of needs: for differing degrees of flexibility in generating the desired information; and for different kinds of content. To some degree, these different needs are represented by different kinds of users, with varying degrees of sophistication in using internet-based "do-it-yourself" tools to generate tables, but we also recognize that any particular user may be able to live with different degrees of flexibility in designing his desired product, according to his purpose. Therefore, we plan to allow users flexibility in obtaining the desired information by providing a product mix which includes final publications; predefined aggregate tabulations (via a data warehouse); flexible tabulations (via a table generator, an example of which can be accessed at **www.cbs.gov.il** - click on **English** and then on the link to the **Social Survey Table Generator**); access to microdata files; and tailor-made tabulations by special order. All of these products and services are already available at the ICBS; we plan to adapt them so they are appropriate for the needs of census data users. All of them will be capable of providing tabulations by census geography, by sub-groups of the population and by subject, individually or in combination. In addition, we hope to develop the capability to prepare analytic reports.

34. One challenge posed by the integrated census is making the results comprehensible to users accustomed to products based on a conventional census methodology. Three aspects of the integrated census require particular attention: geographic detail; linkage to current population estimates; weighted estimates. Unlike in a conventional census, we will enumerate only 20 percent of the dwelling units, and plan to provide census estimates for localities and their statistical subdivisions, down to the level of the statistical area. In a conventional census that enumerates the entire population over all the country's geography, it is possible by using GIS applications to obtain data for any desired geographical area by drawing a border around it and generating the required tabulations for the resulting polygon. The integrated census, however, samples only 20 percent of the geography. While we expect to be able to provide estimates from the 100 percent enumeration for polygons not coterminous with statistical areas, we have not yet determined whether we will be able to do so for the sample enumeration.

35. The ICBS is developing a new procedure for providing current population estimates using data from the Population Registry. Under the existing system, the census provides aggregate population estimates for statistical areas according to combinations of demographic characteristics, and these aggregate estimates are updated on the basis of counts received from the Population Registry on the number of births, deaths and changes of address in each statistical area. The new system is a compromise between aggregate (component) methods and individual (PR) based methods. The 1995 base population was established primarily by aggregate component estimation of the census error. Not everyone in the 1995 base population (the adjusted census population) can be identified with an individual Population Registry record. As a result, the new population base contains "fictitious" records, so that the aggregate base can be treated as though it were a corrected 100 percent count of the population in November, 1995. The 100 percent enumeration in a conventional census provides a list of identified persons that serves as the basis for intercensal updating. The integrated census will not provide such a list, since it will not physically reach each dwelling. Instead, it will provide weights for persons listed in the Population Registry. We have not yet determined how to adapt the new procedure for population estimates to the kind of information provided by the integrated census.

36. A conventional census based on 100 percent enumeration with a short form, and a sample enumeration with a long form, produces two final data files - a demographic file containing the entire census population, and a socio-economic file containing the sample. Whatever error exists in the 100 percent enumeration (overcoverage; undercoverage) is not documented in the final demographic data file, nor is this file viewed as subject to sampling error. The estimates in the final sample socio-economic data file are subject to sampling error. The 100 percent demographic file of the integrated census will also be subject to sampling error, since it is based on information from the 20 percent sample survey. Moreover, the 100 percent demographic file is comprised, essentially, of the updated Population Registry, with each person listed having attached to them a census weight. Aggregate estimates based on the 100 percent file (for geographical areas, population groups, etc.) are, therefore, weighted estimates subject to sampling error. Since users are accustomed to receive population estimates based on a complete enumeration, it will be necessary for us to explain the implications of using estimates based on a sample and assist in their interpretation more than would be required in a conventional census.

IX. FUTURE PLANS FOR THE INTEGRATED CENSUS

37. Israel's 2008 integrated census will serve as the basis for the censuses to follow. The integrated census combines data from administrative registers with data from field enumeration. The cost of the field enumeration is a major part of the cost of any census. Therefore, the long-term goal is to use administrative registers to substitute for the field enumeration and gradually reduce its scope. Planning for future integrated censuses will have to consider a number of issues regarding major components of the integrated census procedures: (1) improving the Population Registry, so that more addresses are accurate; (2) identifying administrative sources of socio-economic data with adequate coverage of the population as an alternative to collecting this information from households; (3) developing statistical tools to provide valid estimates on the basis of partial information; (4) reducing the sample proportion of the field enumeration in order to reduce costs; (5) providing census estimates more frequently.

Section B

X. BASIC CONDITIONS REQUIRED FOR THE CONDUCT OF ISRAEL'S INTEGRATED CENSUS

38. The most important condition required for our integrated census is the existence of a Population Registry having "adequate" address information. I place the word "adequate" in quotation marks because, as I noted above, Israel's Population Registry contains erroneous address information for about 25 percent of the persons listed in it, and an addition 10 percent of them (some half million persons) no longer live in the country. Therefore, we are investing considerable effort in updating the Population Registry in order to improve addresses, using information from other administrative files. As described above, the more accurate the addresses, the fewer persons will fall into the category of PR errors.

39. Updating the Population Registry on the basis of other administrative files requires the existence of these other files and that they contain "better" addresses than those in the PR. These files must be national in scope, preferably in a single national file. They must be computerized, and they must be accessible - that is, the NSO must be able to obtain them from the organization which maintains them, and a solution must be found for possible privacy concerns, both vis-à-vis individual suppliers of files and in the public sphere. The individual records have to be identifiable with the same individual identification number that is used in the Population Registry, so that records can be linked across files. The quality of the identifier must be "good

enough" so that it can serve as the principal key for record linkage (the individual ID number in Israel's Population Registry has a check digit, but not all persons use it, and there may be administrative files in which the check digit is not universally present). In addition to the ID number, there should be in the other files some other information paralleling that in the PR (such as name, date of birth, sex, etc.) that can be used if necessary to verify questionable matches. The quality of that information must be adequate. Programming capabilities are needed to develop the record linkage algorithms. Finally, adequate computer capacity must be available to implement the record linkage procedures.

40. Our procedure requires matching the PR list with the field enumeration list in each EA to identify persons who are registry excess. We search for the registry excess in two stages - first we return to the EA with a list of those who weren't enumerated; second, we try to find the others by telephone. Ideally, the list of persons not enumerated in the EA is generated only after we determine that a person identified as registry excess in one EA was not enumerated in another EA. In other words, we have to match the PR list with the field enumeration for all EAs in the sample. In order for this procedure to work, we have to minimize the time that elapses between the end of the enumeration and the return to the field with the list of registry excess, since people move. A second reason for minimizing this interval is that we don't want to hire and train a new set of enumerators for the overcoverage stage. We want the enumerators who worked in the field survey to continue. But we can't pay them during the interval that they're idle and waiting for us to prepare the tools they need, and we're afraid that some will leave. According to our current timetable, there are almost three weeks between the end of the field enumeration and the start of the field search for the registry excess. We want to reduce the length of this period, and are considering ways of doing so.

41. The principal way in which we minimize the interval between the field enumeration and the field search for registry excess is to insure that all the data from the field enumeration has been captured and is available for the matching as soon as the entire field enumeration is complete. We do so by CAPI enumeration and daily data transmission of the data by the enumerators from their homes via an internet telephone connection. The ICBS introduced this method in its annual Social Survey that began in 2002, and by doing so we are able to prepare a final data file six months after the end of data collection. But the Social Survey is conducted by about fifty interviewers (and needs about 50 laptops). The census will have about 3000 enumerators, and will need 3000 laptops. There is, of course, a trade-off between the size of the enumerator's workload and the duration of field work - the larger the workload, the longer is the time required for field work. On the other hand, a larger workload requires fewer enumerators, and fewer laptops. Israel's population is relatively small (6.8 million); countries with much larger populations might not be able to make the required investment in laptops. Moreover, using CAPI and data transmission from the enumerator's home requires the existence of a universally available national telephone system capable of reliably transmitting the data. Such a system exists in Israel. In theory, all the data from the field enumeration will be in the database the day after the completion of fieldwork.

42. The successful implementation of the telephone search stage which follows the field search for registry excess requires that we are able to obtain information about the registry excess by phone. We hope to reach the person himself, and if we are unable to reach him we hope to contact an immediate family member. Therefore, we need their phone numbers, and we need them linked to the person's PR ID number. The Population Registry doesn't include phone numbers, but the telephone companies (land line and cellular) do have data files which link a person's PR ID number to a phone number. We are able to expand the list of potential phone numbers that we can use to find a person by using the "administrative families" we created and

obtaining the phone numbers for members of these families. We are also able to obtain phone numbers from other administrative data files used to update addresses in the Population Registry (such as household electricity customers). An additional source of phone numbers is information from neighbors obtained by the enumerator in the field overcoverage follow-up about persons he couldn't find. We hope to obtain land-line and cellular phone files which link PR ID numbers with phone numbers in order to facilitate the telephone search stage. One alternative to obtaining these files is to utilize the phone company directory information services, but we don't believe that this will be feasible in view of the anticipated numbers of excess persons we will have to locate.

XI. HUMAN AND FINANCIAL RESOURSE IMPLICATIONS

43. One of the most important reasons for moving to an integrated census was our concern that, in the long run, we would be unable to continue to receive the budgets necessary to conduct a conventional census because of their ever-increasing cost. Israel's previous census, in 1995, cost $11 per person; we estimate that the 2008 census will cost $20 per person. Our assumption in planning the 2008 integrated census is that it will require the same budget that would be necessary to conduct a conventional census, because of the "start-up" costs involved in developing the new procedures and technologies. However, we expect that subsequent integrated censuses will cost less than would a conventional census, because we will be able to utilize the foundation we are laying in preparation for 2008. Whether we will be able to reduce sufficiently the costs of subsequent integrated censuses so that we continue to obtain funding for them is, of course, a question we aren't able to answer today.

44. Since the integrated census enumerates only 20 percent of the population, costs of fieldwork are substantially reduced. On the other hand, it is technologically intensive and requires skilled IT personnel as well as planning staff capable of developing the new census procedures required. We anticipate that the new procedures developed for the first integrated census in 2008 will continue to be applicable to those following, so that the initial "start-up" costs for planning will not recur in subsequent years. We are less certain regarding future IT costs, since it is more than likely that the hardware and applications developed for the 2008 census will not be used in the one following, and the same is probably true for each subsequent census. For example, despite our planned use of CAPI for enumeration in 2008, during the pre-enumeration stage the enumerator canvasses the addresses in his EA and lists the dwellings he encounters in a listing book that he then uses in the enumeration stage. This pre-enumeration listing is conducted with paper and pencil, since we have been unable to identify a portable computer that can be used both for CAPI in the dwelling and conveniently by the enumerator while standing outside of the houses and apartments he is listing. We would prefer to eliminate the listing book and incorporate it into the laptop, but have so far been unable to do so. An alternative would be to provide the enumerator not only with a laptop but with a second hand-held device instead of the listing book. We decided not to examine the feasibility of this option.

45. It is appropriate in considering the human and financial resource implications of the integrated census to discuss in some detail issues that arose in the development of the computer technology required. In addition to the GIS-based applications for creating the national network of EAs and mapping those sampled, there are five major technological systems involved in the operation of the integrated census: the field enumeration; the system that manages the overcoverage operation; the field overcoverage follow-up; the CATI overcoverage follow-up; and the human resources management system. Three different technologies are used: Microsoft .NET [dotNET] (for the field enumeration and the field overcoverage follow-up); Blaise (for the field enumeration, the field overcoverage follow-up and the CATI overcoverage follow-up); and

Magic (for the overcoverage management system, the CATI overcoverage follow-up and human resources management). Why three, and not one? There is a government decision that software development for all government applications be carried out in .NET. However, we have been using Blaise for CAPI and CATI surveys for a number of years, and we believe that Blaise is more appropriate than .NET for computerized questionnaires. On the other hand, Blaise's survey management application does not meet our needs, so we developed our own management application using .NET. And why Magic? Because at the point where it was necessary to develop the overcoverage management system and the system for human resources management the in-house .NET developers were engaged in other tasks, we were unable to obtain approval to recruit additional .NET developers, and in-house Magic developers were available. This added two components to census costs: additional Magic licenses, and technical support services for the applications. Both of these costs could have been avoided were we able to use .NET.

46. A second area in which the technologically-intensive character of the integrated census has affected the allocation of resources is that of data security. The original plan for management of field enumeration envisaged that the enumerator and the crew chief would be in computer contact with one another. In this way the crew chief would be able to track in real time the progress each enumerator was making. In addition, at various points during the field enumeration the enumerator requires authorization from the crew chief in order to continue what he or she is doing. In order for the authorization to be granted, the crew chief must view entries in the enumerator's computer. He could do so via a cellular telephone link, but this would require that both computers be on-line for the duration of the procedure. This was seen as posing an unacceptable risk of unauthorized access to the computers involved, to the census data base in Jerusalem and to other secure areas at the ICBS. Instead, the enumerator transmits to census headquarters in Jerusalem at the end of each working day the information accumulated in his laptop on his progress, and the crew chief can access it the following morning. To do so, however, he must come to the field office, and must physically meet the enumerator every other day, either in the field or in the field office, to carry out his management tasks. This procedure is more cumbersome and less efficient than a direct computer link between enumerator and crew chief would be.

47. On the other hand, the daily transmission of data to census headquarters means that no data on individual respondents is stored in the field offices, unlike the case in a conventional census in which enumerators bring completed questionnaires to the field offices where they are held at least temporarily. That eliminates the need for 24-hour security services at 18 field offices for a period of some months in order to prevent disclosure of confidential information protected by the Statistics Ordinance. Ordinary insurance coverage will compensate us in the event of burglaries; it would not be able to compensate for disclosure of information on respondents.

XII. WHAT DOES THE NEW METHOD GAIN AND LOSE?

48. We have not yet carried out an integrated census, even in a field test, so it is not possible to specify the gains and losses from using it. What we can do is list the expected benefits as well as some of the possible costs that can already be identified. Before doing so, however, it is important to note one clear benefit that is already apparent. The irregularity of censuses in Israel has meant that conducting each new census requires the establishment anew of a census planning team since, in the absence of a permanent census department, census staff dispersed, some to other departments in the ICBS and some to other employment. By the nature of things, the census division tends to separate itself from other departments, which may lead to duplication of effort. In preparing for the 2000 census a decision was made to integrate census planning into the

existing ICBS organizational structure, rather than to create a completely separate operation. While a separate census planning department was established, the implementation of the major components of census planning - IT, geography, field work, statistical methodology, data dissemination, human resources and logistics - is integrated into or coordinated with the ICBS units responsible for these areas. Questionnaire development, and the planning of procedures unique to the integrated census, are carried out by units of the census planning team in a manner similar to the development of methods and data collection instruments by units responsible for particular subject areas.

49. As I have already noted, we anticipate a number of gains as a result of moving from a conventional to an integrated census. Most important, we hope that the reduced cost of an integrated census will enable us to continue to carry out censuses on a regular basis. Second, we will reduce the response burden on the population by eliminating the 100 percent enumeration. Third, we will gain experience in the manipulation of administrative data files that will be applicable to areas in addition to the census. Fourth, we will improve our record linkage capabilities. Fifth, we will identify administrative data files that can provide data on individuals of a quality and scope equivalent to what we obtain from respondents to a survey. Sixth, we hope that what we learn from the integrated census about the quality of information in the Population Registry will encourage steps to improve it. Seventh, using CAPI and Blaise improves the quality of data collected in the field. While this would also be true in a conventional census, it would be too expensive to equip each enumerator in such a census with a laptop. Therefore, an additional benefit of moving to an integrated census is improving the quality of the demographic and socio-economic data we obtain. Eighth, linking the GIS/mapping process and the allocation of PR "administrative households" among EAs enables computerized quality control of field enumeration, in particular errors resulting from EA "boundary transgressions" - incorrectly enumerating households located beyond the EA boundary.

50. We are able to anticipate a number of potential problems that will result from the move to an integrated census, though at this point it is difficult to evaluate how serious they will be or what consequences they will have. First, as I noted above, we don't plan at present to provide estimates for data based on the sample enumeration for geographical areas that are not multiples of localities or of statistical areas within localities. In a conventional census that enumerates everyone and samples systematically across all geography it is possible to provide estimates for any geographical polygon. To the degree that we are successful in anchoring the persons listed in the PR to their correct addresses, we will be able to provide such estimates for the 100% demographic count. We don't know at present whether we will be able to do so for the sample socioeconomic data. Second, because the socioeconomic sample data is based on a cluster sample of dwellings and not a systematic sample, estimates based on it will have a larger variance. For the same reason, our estimates of household characteristics will be poorer than our estimates of individual characteristics. Third, in a conventional census based on a complete enumeration we are able to provide preliminary estimates of the number of persons and households by geographic area immediately upon the conclusion of field work, using data from the enumerators' listing books. We don't know whether we will be able to do so in 2008. Fourth, it is not yet clear how or whether we will employ the estimates from the integrated census as the basis for the new procedure we are developing for providing current population estimates. Fifth, the census file is used as a sampling frame for post-censal surveys. If the 100 percent file is the basis for the frame, the addresses it contains will be less accurate than would have been the case in a conventional census, since only 20 percent of the households will actually have been enumerated at a verified address (for 80 percent the address is based on probabilities). If the 20 percent sample is the basis for the frame, the sampling procedure will have to take into consideration that, unlike in a conventional census, there is variation in the sampling fraction

across geographies, and that the 20 percent sample for the country as a whole does not necessarily hold for any particular area. Sixth, the conception underlying the integrated census is harder to explain than that of a conventional census since we haven't actually enumerated most of the population, and we may have problems explaining to users and to the public what the numbers mean.

XIII. HOW IS THE TOTAL COVERAGE AFFECTED? ARE THERE PROBLEMS OF BIASED ENUMERATION?

51. "Total coverage" in a conventional census refers both to the success of census staff in enumerating every person in the census population, and to the result of the procedure intended to insure that the entire area of the country defined as included within the census boundaries is canvassed. "Total coverage" in the integrated census has a different meaning. The census population is, by definition, the total of the persons listed in the Population Registry and actually present in Israel (or absent for less than a year), plus the number of persons in defined categories who are not listed in the PR but who have been resident in Israel for at least a year. We assume that everyone who is supposed to be listed in the PR is actually listed there.

52. The only persons "missed" by the PR are those who aren't supposed to be there - persons without an Israeli ID number. They fall into the following main categories: diplomats (who are not part of the census population); foreign workers (some of whom, such as caretakers for the elderly, live in census households; others, in construction or agriculture, may live in employer-provided accommodations or haphazardly at work sites); students from abroad studying either in institutions of higher education or in Jewish religious institutions and live either in households or in accommodations provided by the school. Persons in the latter two categories may be present legally or illegally, but the census cares only about the duration of their residence in the country. We can know of their existence in two ways - by coming across them during the field enumeration of sampled EAs, or by utilizing administrative registers that list them. Foreign workers who entered legally are listed in such a register, as are students who entered with student visas. In principle, we could employ the same method to estimate their numbers as we do with the population listed in the PR: match the results of the field enumeration with the registers and look for the persons in the category of registry excess. We have not yet decided what procedures we will employ to include these groups in our estimates.

53. The integrated census will provide estimates based on a combination of enumerating persons living in households in a sample of the EAs and 100 percent of persons in special populations such as residents of communal quarters, residents of kibbutzim (collective settlements) and Bedouin living outside of localities in southern Israel. In localities that are divided into statistical areas we will sample 20 percent of the EAs in each statistical area. We have not yet decided whether we will sample 20 percent of the EAs in each locality too small to be divided into statistical areas, or group such localities into clusters having similar demographic characteristics and sample from within the cluster. Neither method should result in biased estimates.

54. We have not yet decided whether the census will provide an estimate of the number of foreign workers in Israel. The census questionnaire will obtain information on persons lacking a PR ID number regarding how long they have been in the country in order to determine whether they should be included in the census population, and we also intend to distinguish, for persons present less than one year, those who have been in Israel more than ninety days and those who have been present for a shorter time. Many foreign workers live in irregular accommodations, at construction sites, in industrial areas and at other non-residential locations. The network of EAs

constructed for the census is based on the number of persons listed at Population Registry addresses, but many foreign workers don't live in structures whose addresses appear in the PR. Thus, making the enumeration of foreign workers a goal of the census would require expanding the EA network to non-residential areas. Moreover, since many foreign workers are in Israel illegally, they would be very difficult to enumerate even if we succeed in locating them. For these reasons, the census unlikely to enumerate many foreign workers, even those present legally. This is a potential source of coverage bias.

55. Coverage bias could also result from incorrect construction of the weighting groups. As described above (Par. 17), our goal is to construct weighting groups that are homogenous with respect to the probabilities of their members being allocated to one of the four outcome categories, across all localities and statistical areas. Departures from homogeneity may introduce bias.

XIV. HOW IS THE QUALITY OF THE DATA AFFECTED?

56. I will discuss separately the 100 percent demographic data and the socio-economic data from the 20 percent sample. In a conventional census, two major sources of error affect the 100 percent demographic estimates: errors of overcoverage and undercoverage, and response error. Estimates of the extent of such errors are provided by post-enumeration surveys. The 100 percent demographic estimates based on the integrated census are also subject to errors of overcoverage and undercoverage. Some of these potential errors are mitigated by the search for registry excess, which will potentially identify some of the EA field undercoverage, because it shows up as registry excess in another sampled EA. Similarly, some field overcoverage in an sampled EA (incorrectly deciding that a person should be enumerated at a particular address) will show up as registry excess in another sampled EA and can be corrected during the overcoverage search stage.

57. Other potential errors in the 100 percent demographic estimates stem from the procedures that create the updated PR file. Recall that we begin with an inaccurate PR file that contains persons who are not included in the census population whose addresses are incorrect for about 25 percent of the persons who are included in the census population. Moreover, the PR file that is matched with the results of the field enumeration to identify registry excess, though generated to reflect the content of the PR on census night, may lack information on some births and deaths that occurred beforehand. That last shortcoming is easily overcome by corrections based on receipt of a subsequent PR update file. We try to correct the first two shortcomings by using other administrative files to update the population and the addresses. This process introduces two kinds of errors: those resulting from incorrect record linkage between the PR and the external administrative files and those resulting from incorrect decisions made by algorithms (such as determining whether a person not present in the country should be defined as an emigrant, and hence not part of the census population, or as someone temporarily absent, and hence to be counted; or those involved in choosing the "right" address among alternatives in different files).

58. A third source of potential error in the 100 percent demographic estimates from the integrated census stems from the assignment of weights to persons in the PR in non-sampled EAs based on the outcome probabilities computed for those listed in the PR in the sampled EAs. The quality of these weights depends on three factors: the degree to which the sampled EAs are representative of all the EAs in a statistical area; the degree to which the outcomes for the persons in the sampled EAs are representative of the outcomes for the population of the statistical area; and the quality of the information which is obtained in the overcoverage search stage and which determines the outcomes. Errors in this procedure will lead to errors in the geographical placement of the overcoverage

59. The final 20 percent sample file reflects the results of the field enumeration and the allocation to addresses of registry excess in the overcoverage search stage. This allocation depends on our success in locating the registry excess and on the quality of the information obtained in the overcoverage search stage, which, as in the case of the 100 percent demographic file, affects the geographical placement of the overcoverage. In addition to the coverage errors in the sample file the data it contains, obtained in the survey, are subject to response error. As noted above, using CAPI and Blaise improves the quality of data collected in the field.

XV. WILL THE NEW METHOD CHANGE THE REFERENCE POPULATIONS (for example, de facto/de jure) AND WILL THIS AFFECT THE COMPARABILITY WITH PREVIOUS CENSUSES?

60. The new method does not change the reference population, which comprises all persons with a PR ID number unless they have been abroad continuously for a year or more, and persons without a PR ID number who have been present continuously in Israel for at least one year. In that sense, comparability with previous censuses should not be affected. Since the population estimates from the integrated census are not based on a complete enumeration but on statistical modeling, comparability between the results of the integrated census and a complete enumeration depends on the quality of the models and their implementation. On the other hand, were we conducting a conventional census, comparability with previous censuses would be affected by differences in the coverage of populations and geographical areas from one census to the next, and the methods for imputing missing information. (Using CAPI for the field enumeration allows us more flexibility than is possible with a paper-and-pencil questionnaire to identify persons who are not members of the reference population but for whom we are interested in obtaining information - those present in the country for more than three months but less than one year.)

*****