Major Post Censal Redesign of Household Sample Surveys in the United States

by Preston Jay Waite[1]
U.S. Bureau of the Census

Acknowledgements

I wish to acknowledge the work of many staff members in the Statistical Methods Division of the U.S. Bureau of the Census in the preparation of materials for this paper. Special thanks to Dr. Charles Alexander and Gary Shapiro for their contributions to the editing and review and to Pat Curran for her work in preparing the manuscript.

Post Census Surveys (PCS) are utilized in a variety of ways in the United States. A large sample (one household in six) is imbedded in the Census collection itself. This sample allows for the collection of detailed housing and persons information not covered by a complete count. We also use sampling for the measurement of undercoverage by selecting a large post-enumeration survey (PES) of approximately 150,000 households. The results of this survey are then matched to the census enumeration to measure the extent and characteristics of census undercoverage. Major census follow on surveys of Residential Finance and of scientists and engineers are conducted immediately following the census. Frames for these surveys are constructed by screening units with particular characteristics from census questionnaires.

All of these survey collections are major operations and complete papers suitable for this conference could have been produced for each of them. I would like to focus my remarks today, however, on an additional use of the census that being for a frame for selection of the major household surveys conducted by the United States Government.

This paper will discuss our ongoing plans to redesign our current household surveys based on the 1990 census. I will discuss our general methodology and the challenges we face by trying to simultaneously select several surveys simultaneously. I will also mention briefly some of the planned uses of new technologies in the reselection of our survey samples.

Using the Census As A Frame For Continuing Household Survey

The United States Decennial Census Address Lists are used as a sampling frame for many of the Government's major continuing household surveys. The principal household surveys using the census as a frame are:

1. The Current Population Survey, (sponsored jointly by the Labor and Commerce Departments; the basic labor force survey).

2. The Consumer Expenditure Surveys, (sponsored by the Labor Department; used as input to the Consumer Price Index).

3. The Current Point of Purchase Survey, (sponsored by the Labor Department; consumers are interviewed to generate a frame of retail outlets for measuring prices for the Consumer Price Index).

4. The Survey of Income and Program Participation, (sponsored by the Commerce Department; a longitudinal survey which follows persons every four months for two-and-one-half years to collect information on income dynamics and use of government transfer payments programs).

5. The National Crime Survey (sponsored by the Justice Department; collects information from victims of crime).

6. The American Housing Survey (sponsored by the Department of Housing and Urban Development; a biennial longitudinal survey of housing which updates the census data for sample units, while adding in new construction).

The census address lists are the primary source of the sample for all of these surveys. These lists give us a frame for the United States as of census time 1990. Since these are continuing surveys, the sampling frame derived from the census must be kept up to date between censuses. This is done mainly by sampling building permits, which are required for new construction in most parts of the country. Permits are listed and sampled every month from selected building permit offices. Where such permits are not required by local governments, new construction is represented through area sampling. In the area sample, a list of all the addresses for selected areas on the map is made by the field representative. Area sampling is also used for both old (existing in 1990) and new construction in some mostly rural areas where permits are not required for new construction and for areas where the census addresses are hard to locate.

The census address list is an inexpensive source of sample, compared to an area sampling approach, and it

gives more complete coverage of the population than any other available list of addresses. But even with the census list we find coverage to be a problem. Potential coverage problems can be of two types; coverage of households and coverage of persons within households.

Evaluation of the coverage of households in the census shows that at most a one to two percent overall undercoverage at the time of the census, although undercoverage for households of minority races is known to be substantially worse than for the population as a whole. Coverage of households by the continually updated census frame is more difficult to measure, but estimates of missed households range from about one to five percent. Estimates of missed persons are more reliable, since the survey estimates of the number of persons may be compared to updated census estimates. This comparison shows on the order of a 10 percent undercoverage of persons, with the worst coverage for young males. There is evidence that for young males most of this is due to failure to obtain complete lists of household members, rather than to missing households.

The updated census estimates of persons by age, race, and sex are produced by inflating or deflating the census counts for births, deaths, immigration and emigration. Most of the survey estimates are calculated using post-stratification to bring the final survey estimates of persons into agreement with the updated census estimates.

The surveys using the census as a frame are all conducted by the Bureau of the Census, although the data may be analyzed and published by other government agencies or research organizations. By law, no one outside the Census Bureau may have access to the actual census addresses, nor to information which would permit the identification of any sample household. This places limits on the amount of detail which can be included on data files intended for public use. It also means that only sworn Census Bureau agents may contact the sample households for any survey which uses the census as a frame. To avoid these restrictions, another major household survey conducted by the Census Bureau, the National Health Interview Survey, uses only an area sample. The operations for this survey are coordinated to some extent with the other household surveys, and it is redesigned in conjunction with them, but the frame for this survey is created independent of the census.

Having a single field staff conduct the interviews for all survey is extremely cost-effective. The administrative and office costs can be shared. Also, in many cases, the same field representative can conduct interviews for several surveys, since the surveys take place at different times of the month. This sharing of field representatives reduces the number of field representatives who have to be recruited and trained. Detailed coordination of the sampling operations also saves effort and money. The lists of building permits used to keep the frame up to date can be shared among the surveys. In the area frame, the later surveys in a particular map area can make use of the lists made for the earlier surveys. The cost savings from sharing listings are substantial. Since the sample addresses for the different surveys are kept close together whenever possible.

Redesigning the Sample After Each Census

The census address list was first used as a sampling frame for the ongoing household surveys following the 1960 census. In theory, this frame could have been kept up to date perpetually by adding new construction from building permits and area listings. In reality, the sample was reselected after the 1970 census and again after the 1980 census. The sample will be reselected after the 1990 census. One reason for reselection is the likelihood that in spite of our best efforts, continued updating of the old frame inevitably will lead to a gradual deterioration of coverage. Additionally, as time goes on, a greater proportion of the sample would come from the more expensive permit frame.

A basic reason for redesigning the sample after each census is to use information from the new census in improving the design. The census information collected for each household includes household size, race of the occupants, whether the unit is a farm, whether the dwelling is rented or owned, and the rent or value of the dwelling. A sample of about one-sixth of the households receive a "long form" during the census, which asks for many additional details about the dwelling and its occupants, including income and labor force status.

All this information is used in the redesign to restratify the primary sampling units, so as to reflect changes which occur between the censuses. The economic characteristics of many metropolitan areas have changed in recent decades, and there has been a shift of population to some formerly less developed parts of the "Sunbelt" in the southern portion of the United States. Areas which were similar 10 to 20 years ago may be very different today. Taking these changes into account results in more efficient and reliable samples.

Prior to the 1980 redesign, all the surveys used the same general purpose sample design. This general purpose design was modified somewhat for the Current Population Survey (CPS) to improve the measurement of labor force data. The other surveys had smaller sample sizes than the CPS, but used the same PSUs, the same cluster size, and the same within-PSU stratification. Indeed, the other surveys were merely allocated a portion of the extra "reserve" sample which was selected for CPS along with its regular sample. As the importance of the other surveys grew, greater attention was paid to their sample designs. Following the 1980 census, the surveys were redesigned individually, in an attempt to optimize them based on their separate objectives, rather than using a common design modified only for measuring labor force characteristics.

The expenditure and housing surveys sort units based on census characteristics which are highly correlated with the variables measured by the survey. The expenditure surveys concentrate on rent or value of housing, which is asked of all units in the census. The housing survey takes a subsample of census "long form" households, for

which detailed housing characteristics are available. The other five surveys do not sort individual households using census characteristics, either because the relevant questions are not asked in the census, or because the relevant variables are not stable over time and the benefits of sorting would quickly dissipate. Another reason for not sorting separately for all surveys is that some surveys interview clusters of adjacent households to reduce travel costs. Some sorting, such as separating urban and rural areas within each county, is used for all the surveys.

The CPS is still our largest survey and as such still has some effect on the others. The CPS is the only survey that attempts to measure data for states as well as for the United States as a whole. In 1980, the CPS sample was selected as 51 independent state samples, one from each of the 50 states and the District of Columbia. This was necessary because there was a reliability requirement for the unemployment estimates for each state. The need for reliable state data was in response to the allocation of Federal funds determined in part by the estimated state unemployment rates. The other surveys use sample designs aimed at making national estimates, and therefore their primary sampling units may cross state lines.

Although each survey now has its own stratification of primary sampling units, steps were taken in the 1980 redesign to maximize the selection of common sample areas across surveys. This will be done again in the 1990 redesign. This allows field representatives to be shared, and allows the permit and area samples to be better coordinated. The largest metropolitan areas are automatically in sample for all the surveys. Several of the surveys select their sample PSUs as a subset of the CPS PSUs. The crimes survey used a "maximum overlap" method, in which each PSU was given the appropriate unconditional probability of selection, while the expected amount of overlap with the CPS was maximized. (This maximum overlap sample selection is still commonly called "keyfitzing" after Nathan Keyfitz, who developed one of the original methods). A different mathematical method is now used, but the objectives are the same as Keyfitz addressed. The same maximizing technique will be used to select the new 1990 CPS PSUs from their new strata while maximizing the expected overlap with the old CPS sample PSUs. This was done to reduce the need to train new interviewers in some areas while having to lay off interviewers in old areas.

Effect On New Technologies

The last 10 years have seen increasing automation of census data products and survey interviewing techniques. This affects the sample redesign directly because automation offers potential new efficiencies in the sampling operations, and indirectly because changes in design may be needed to make the most efficient use of the new interviewing technologies.

The most dramatic change in interviewing techniques has been the introduction of computer-assisted interviewing, combined with increased use of telephones by the interviewers in the field. The Census Bureau has recently opened a centralized computer-assisted telephone interviewing (CATI) facility, located in Hagerstown, Maryland. Interviewers at the facility work from a computer terminal which automatically selects the next case for interview, schedules callbacks, and displays the questionnaire on the terminal's screen. The computer program determines the path through the questionnaire based on the answers which are entered, checking the responses for consistency as it goes. We expect this system to improve the control of data quality, both because of the control provided by the computer and because interviewers can be closely monitored by the supervisors at the centralized facility. The system eliminates labor-intensive data entry and some steps in processing, which are needed for paper questionnaires.

CATI has been tested successfully for the National Crime Survey and is being tested for the CPS. We expect the CATI methodology to be in full-scale use in the early 1990's. Its use for household surveys has so far been restricted mainly to follow up interviews for panel surveys. An address sample is still used and the first visit to a household is made in person. Even when full CATI is being used, there will still be a need for field interviewers. Not every housing unit has a telephone and some that do request a face-to-face interview. Testing the use of computer-assisted personal interviewing to complement CATI is now underway and is also well underway.

Some testing using samples based on randomly selecting telephone numbers has been done by the Census Bureau. However, the response rates in these tests were much lower than when the first visit was made in person. Because of this, along with the need to represent households without telephones, a purely telephone sampling approach is no longer considered for most of the household surveys. An exception is the Current Point of Purchase Survey (CPP), which is just completing a test of a combined telephone-list/random-digit-dialing approach, with promising results. This could eventually remove CPP from the list of surveys using the census as a frame. We plan to select some CPP sample in the 1990 redesign as a backup strategy in the event that the random-digit dialing approach proves unsuccessful.

One aspect of the 1990 sample redesign will be to modify the sample designs to make more efficient use of centralized CATI. CATI removes some of the follow up interviewing from the dispersed field representatives. This means that the field interviewers will be underutilized unless they are given greater workloads initially. Thus, for a constant total budget, the optimal design using centralized CATI will have fewer PSUs, with a larger initial workload in each PSU. The increased use of telephoning also reduces the relative importance of travel costs, which may reduce the optimal cluster size for those surveys which select clusters of households.

The increased use of computers in the 1990 decennial census will make it easier than ever before to use census data in constructing a sample frame. The most important innovation is a geographic database known as the Topologically Integrated Geographic Encoding and

Referencing (TIGER) system. This computerized system (developed jointly by the Census Bureau and the U.S. Geological Service) will produce a map of any city block or comparable rural "block," with roads and natural features correctly represented. Addresses from the census will be linked to the correct block on the map and in most instances the location within the block will be indicated. The TIGER maps have the potential to revolutionize the area sampling operations. In the past the development of maps has been a particular problem and staff members have struggled with maps and information of inconsistent quality from a variety of sources. The computer generated maps will also simplify locating those new units whose building permits have been selected. The TIGER maps and data (without detailed census address information) will be available to the general public and will be useful for area sampling by survey organizations outside the Census Bureau.

Another 1990 census product which will facilitate using the census as a sampling frame is the automated Address Control File. This contains a record for each census address, with basic information about the housing unit at that address. For about 95 percent of the records, the actual address will be included as text on the file. In previous censuses, the addresses could only be obtained by going to the handwritten register completed by the census enumerator, which necessitated an expensive address keying operation before the surveys could use these addresses.

Computer technology was also used to advantage in implementing the mathematical methods for stratifying and selecting PSUs in the 1980 redesign. Similar methods will be used in 1990. The CPS strata before the 1980 redesign were formed by writing key PSU characteristics on 3x5 index cards and grouping the cards manually to form intuitively homogeneous strata of roughly equal stratum population. For the 1980 redesign, a multivariable clustering algorithm was modified to form strata so as to minimize a measure of total variance for a set of specified variables, subject to constraints on the stratum population. Also, for the 1980 redesign, an improved "maximum overlap" method was developed, which selected a probability sample of PSUs while maximizing the overlap with some other survey's selected areas. This method used a linear programming algorithm to maximize the expected overlap, subject to constraints on the probabilities. This gave a greater percentage of common PSUs than methods used previously.

Coordinating Sampling With Different Sample Designs

A central theme of the 1990 redesign research is to better coordinate the sample selection operatiions for the different surveys. As I have described already, in the 1980 redesign we allowed different surveys to use different PSUs and different ways of sorting, stratifying, and selecting households within the PSUs. At the same time, every effort was made to keep the surveys' sample units close together to save on the cost of keying addresses, listing for the area sample, and sampling building permits.

This task of linking different sample designs turned out to be quite complicated. An example is the coordination of listings for the area sample. If surveys are to share each other's lists of households in sample blocks, then it is necessary to keep a cross-referenced index so that each survey can find out who else has previously made a list of the block and where that list is being kept. This sort of thing was much easier in the 1970 redesign where there was only the one CPS design, so that it was only necessary to find out whether the block had been listed for the previous CPS sample.

Extensive record-keeping is also needed to avoid duplicate selection of the same address by different surveys. United States Government statistical policy, as set down by the Office of Management and Budget, is that a single address should not be included in more than one Census Bureau survey. With different surveys selecting sample from the same universe, using different method, it was not easy to avoid such duplication. Particular problems were in the Health Interview Survey, which used an area sample where the other surveys were using list sampling, and the housing survey, which selected specific long-form units where the other surveys were using area sampling.

All this complex record-keeping and cross-referencing is amply justified by the large savings from coordinating the survey operations. However, the complexity becomes a liability if at any time between redesigns, one of the surveys has its sample reduced, expanded, or has a change in the scheduled interview dates. When such changes are made, all the references to the changed survey anywhere in the reference system must be checked and updated, to avoid duplication and other operational problems for the other surveys. Because the system was not designed with updating in mind, this causes even small changes in a survey's sample to be time-consuming and expensive, even when the changes can be made by computer.

The sampling of building permits is especially inflexible. Permits are sampled every month as they are issued by the permit offices. (There are over 10,000 of these offices throughout the country.) Many of the offices destroy their old records after a few years, so it is impossible to go back and select more permits.

To try to simplify the record-keeping in the 1990 redesign, we will closely examine the details of all the clerical and computer procedures, and standardize these procedures whenever possible. Some of the major research issues concern whether specific surveys would incur significantly higher field costs or higher variances by simplifying on certain operational details in order to standardize their procedures with the other surveys. In the 1980 redesign, four separate sets of computer programs were used to select the sample for the seven surveys. (Some surveys were able to share programs). These programs were developed separately and there were minor differences in definitions and data formats

which turned out to be major barriers to coordination. In the 1990 redesign, the plan is to use one integrated set of computer programs for the entire sampling operation. In developing these programs, our programmers propose to use a Computer-Assisted Software Engineering approach. Besides providing a common logical framework for the computer algorithms, the computer-assisted planning produces a common data dictionary for all the programs. This will enforce standardization of definitions across the surveys.

Selecting several surveys from the same frame can add complications to the mathematics of sample selection. A basic example is that if a survey removes units from the universe with probability proportional to size, then the remaining universe tends to under-represent large units. Such concerns need to be kept in mind as the sample selection methods are designed. A goal of the 1990 redesign is to leave a "clean" universe, so that future surveys can be selected from the census address lists, building permit offices, and the area frame without having to make special adjustments because of the sample of units which has been "removed" by the redesigned household surveys.

One final challenge in coordinating sample selection for multiple surveys is getting agreement on a time for selecting the sample. It is most economical to do the bulk of the sample selection work at the same time for all surveys. However, this requires all the different sponsoring organizations to complete their research on the new sample design in time. Some agencies prefer to have their redesigned sample introduced as soon as possible after the 1990 census, to take advantage of the new design. Others would benefit by having more time to use the 1990 census data in research on special topics affecting their survey, before deciding how to design their survey.

"As soon as possible" after the 1990 census turns out to be nearly four years later; the 1990 redesign sample will start being introduced in April 1994. Part of this lag is due to the census processing; the last 1990 census data file used in sample selection becomes available about 18 months after the official April 1, 1990 census day. Once the design has been specified, computer processing to prepare materials for the clerical work takes about 12 months, and the various clerical activities and related processing take about 9 months. This leaves about 9 months to use the1990 census data to specify the design, including selecting PSUs, deciding how to stratify units within PSUs, and deciding on the sample size at each stage of selection. Obviously most of the basic research, planning, and software design has to take place prior to the availability of the 1990 census data. ❑

[1]Presented at the IFDO/IASSIST 89 Conference held in Jerusalem, Israel, May 15-18, 1989.

NOTICE TO IASSIST READERS

To all **IASSIST** members

We are trying to collect as many photos as possible taken at **IASSIST** conferences or other official functions. If you have pictures please send a copy to:

Sue Gavrel
129 Blackburn Ave
Ottawa, Canada K1N 8A6

We hope to have an album or two for the Edmonton Conference