

1990 HOUSEHOLD SAMPLE REDESIGN

Preston Jay Waite, United States of the Bureau Census
Demographic Statistical Methods Division, Room 3705, Building 3, Washington, D.C. 20233-4700

Keywords: Oversampling, unduplication, coordinated selection

INTRODUCTION

The Household Sample Redesign is a complex project. It requires a substantial effort over several years with mostly new and inexperienced staff. There are truly many technical and statistical issues that are required to be merged and coordinated in order to complete the sample redesign. Like many large undertakings, however, one of the major challenges of the redesign is managerial and operational. Simultaneously coordinating the survey design needs of eight major household surveys and designing a common system to handle the operational consequences of their needs is a daunting task. Because of field cost implications there is a high degree of overlap and coordination of these surveys, but a strict duplication of sample households across surveys is prohibited by a directive from the Office of Management and Budget (OMB). Getting survey sponsors to agree on a common system and on timing for activities that can be incorporated into a unified schedule is to say the least challenging. Add to this the desire on the part of the Census Bureau to develop an automated data collection capability unevenly across surveys and with an evolving time schedule and you have the makings of a real interesting job. This paper will describe briefly the surveys involved in the sample redesign. Why we need to do a redesign and a brief overview of some major areas that we addressed.

The sample redesign, while not yet completed, has been an unquestioned success. We have consistently been ahead of schedule and under budget. The continual crisis mode so often associated with this type of activity has been virtually nonexistent. The work was well thought out, planned and executed with a

minimum of disruption. From the very beginning, we employed a project development plan (PDP) approach. This management approach assumed: 1) an overview of the system, explicitly including all assumptions, should be prepared and approved at the start of the project; 2) planning should be top-down, with an emphasis on filling in the least certain areas of the overview before implementation of any part of the system begins; 3) any area of the division which will ever be involved in implementing a particular system should be brought into planning from the beginning. This was reflected in the organization and reporting structure of the redesign workgroups. Based on our redesign experience, I would say that PDP works.

In comparison with the more traditional approach "begin in haste, redo in leisure," the PDP approach is clearly superior. The start is slower, but the finish is much faster and the ride is more pleasant.

I would like to take credit for this operational success of the 1990 Sample Redesign, but I'm afraid I can't. This task while truly remarkable has been accomplished by a relatively inexperienced staff with fresh ideas and a desire to get the job done. The success of the project is truly theirs. They accomplished a big job without fanfare and without crisis management, not because of senior division management, but in many cases in spite of it.

The final completion of the sample redesign is probably a couple of years away, but the course is set and success is assured.

WHAT SURVEYS ARE INCLUDED IN THE SAMPLE REDESIGN

The redesign of the samples for all of our major household surveys is done once each decade following our decennial census. This is a coordinated effort involving the simultaneous reselection of samples and modification of sampling and field collection procedures for

eight major ongoing household surveys. The surveys covered by the sample redesign include:

1. The Current Population Survey (CPS) (conducted under joint sponsorship of the Bureau of the Census and The Bureau of Labor Statistics)
2. The Survey of Income and Program Participation (SIPP) (sponsored by the Bureau of the Census)
3. The National Health Interview Survey (NHIS) (sponsored by the National Center for Health Statistics)
4. The National Crime Victimization Survey (NCVS) (sponsored by the Bureau of Justice Statistics)
5. The Consumer Expenditure Survey (Quarterly and Diary) (CES) (sponsored by the Bureau of Labor Statistics)
6. The Consumer Point of Purchase Survey (CPP) (sponsored by the Bureau of Labor Statistics)
7. The American Housing Survey National Sample (AHS-National) (sponsored by the Department of Housing and Urban Development)
8. The American Housing Survey Metropolitan Sample (AHS-MS) (sponsored by the Department of Housing and Urban Development)

OBJECTIVES OF THE POST 1990 REDESIGN

For the post-1990 redesign, we embarked on a more limited research program choosing to rely heavily on the work done during the post-1980 redesign. Our efforts in the 1990 redesign were directed more to the operational aspects of the sampling and to how we could efficiently coordinate the needs of the surveys. Because of this operational emphasis, we felt that it was critical that we involve operations and the Electronic Data Processing (EDP) personnel much more extensively the very early planning for the sample redesign. This has proved to be extremely beneficial. As problems with details of the various designs arise, staff from all areas of the division have common understanding not only of what is

needed, but the likely effect of alternative solutions on other aspects of the sampling systems. The usual finger pointing and cross blaming between staff of different disciplines has been virtually eliminated and a real team effort has developed. I believe that this more than any one single thing has been responsible for the work on the sample redesign proceeding consistently ahead of schedule and under budget. For those who have been involved in earlier sample redesigns, this orderly completion of tasks is indeed a refreshing change.

Because of our desire to complement rather than duplicate the emphasis of the 1980 redesign and partly because of our perception that budget and staff resources would be limited, we embarked on an operations focused redesign as follows.

A. Reselect sample using 1990 census materials, modifying procedures as needed to take advantage of any significant simplifications and cost savings made possible by increased census automation.

B. Reevaluate design parameters and sampling methods to ensure that there were no obvious changes necessary because of the emerging computer-assisted instruction (CAI) methods, changed sponsor need and interests, and so forth. The research on design parameters relied heavily on what was learned in the 1980 research, but new customer needs did require some new research.

C. Redefining, restratification and reselection of PSUs using 1990 census data to form the strata. Selection of stratification variables and decisions about number of PSUs across designs was primarily based on the 1980 research. One area where new research was undertaken involved expanding on maximization of PSU overlap work. This methodology was expanded to include two PSU per stratum designs. The selection of PSU for all surveys was coordinated using a basic building block approach.

D. Investigate a limited number of the most promising methodological innovations.

1. Improvements in unduplication of sample and coordination of surveys as well as the

operational systems needed to more efficiently expand and reduce samples over time.

2. We identified areas that were most expensive for our ongoing sampling operations and examined possibilities of reducing costs through simplification of our procedures.

3. We sought to decrease the sampling rate in our "expensive" permit and area frames as well as in our high travel rural areas.

4. We researched improvements in the within-PSU sampling such as considering equal person cluster, noncompact segments and changes in sorting and stratifying units and addresses within the PSU.

5. Several sponsors were interested in expanding or initiating oversampling procedures for special populations. These issues involve whether or not to use screening, how to screen, how to use the census data to oversample groups and at what level should we try to oversample.

BASIC SAMPLING STRATEGY

There are many aspects of this redesign that offered special challenges to the survey designs. I will briefly discuss only three here.

1. Unduplication

2. Unit segments replace address segments

3. Oversampling

A. Unduplication

One of the unique aspects of the combined household sample redesign is that we try to unduplicate across surveys. This strange practice rises out of the Office of Management and Budget (OMB) directive that, in the spirit of burden reduction, states that we should avoid including the same household in multiple surveys. We at the Census Bureau have taken this advice and expanded it into a religious experience. We have committed ourselves to doing everything humanly possible to avoid any household being in more than one of our major household surveys during the entire decade covered by the sample redesign.

This objective has significantly complicated the overall redesign process. We set out to ensure nearly total unduplication for the decade through an automated system with minimum or no clerical intervention. To do this, we needed to maximize the coordination

and consistency of the sampling process across all surveys. This meant that within a geographic area, usually a block, and within a given frame, all surveys needed to use the same sampling process. We spent a tremendous amount of resources and intellectual capital in coming as close as possible to achieving this goal. In retrospect, I don't think that it was worth it. One of my major objectives for future redesigns is to relax this onerous and extremely expensive feature of our sampling systems.

Samples for the redesign are selected from four nonoverlapping frames. Each frame is sampled separately and with its own sampling procedure, but within a given frame the sampling process needs to be the same for all surveys.

1. Unit Frame

This is the file of 1990 decennial census serial numbers within the set of blocks NOT designated as area sample blocks. The samples for all surveys conducted under the Authority of Title 13, we selected primarily from the unit frame. For this frame, samples for surveys were selected sequentially to avoid selecting the same census serial numbers. The serial numbers on this frame may be sorted differently within a block for different surveys, but once a serial number is selected by one survey, it is flagged to avoid its selection by a subsequent survey. When multiple surveys select sample from the same blocks, the sampling intervals within that block were adjusted for subsequent surveys in order to achieve the desired sample size.

2. Area Frame

The area frame consists of a file of blocks that were included in the census, but were removed from the unit frame universe prior to unit frame sampling. A census block is selected to be in the area frame, if any of the following conditions exist.

a. The block is in an area not covered by building permits. We use building permit data to get a sample of units constructed since the last census. For blocks not covered by permit issuing offices, we need to do a listing and determine year built to get a sample of new construction.

b. The block is in an area covered by building permits, but there is a large number of incomplete addresses in the block, making census serial number sampling problematic.

c. The block is selected as part of the all area HIS sample and therefore, for unduplication purposes must be treated as all area for other surveys as well.

For this frame, we select measures of expected size of four addresses. Specific addresses are associated with a sample measure through the listing of addresses in blocks. Similar to the unit frame, for unduplication purposes it is necessary to remove the first surveys selected measures from the universe and adjust the sampling intervals of the subsequent surveys.

3. New Construction Frame

We create a list of addresses of structures built since the last census by going to building permit offices in our selected PSUs and copying the addresses from the permits they have issued. The selection of sample is automated and ensures unduplication. The selected measures have to be removed and sample intervals adjusted for subsequent surveys for this frame as well.

4. Group Quarters

The final frame for selection is the group quarters frame. This frame consists of all the group quarters addresses reported in the census. These are school dormitories, hospitals, military bases, hotels, motels, nursing homes, and so forth. Field staff list the noninstitutional living quarters within each group quarter selected for sample. Special procedures are developed for removing measures of group quarters selected by multiple surveys.

To ensure unduplication, surveys in the same block or permit must use the same listing of addresses. (Note, the sharing of listings is also necessary in the unit frame to ensure unduplication when a specific unit cannot be located in a multiunit address and the address is listed to identify sample units.)

As long as each survey classifies each block into the same frame, unduplication can be ensured using the above procedures. This is theoretically correct, but in practice is difficult

to carry out. For example, NHIS has an all-area design, while the other surveys have multiple frame designs. The same block may be initially classified into different frames by NHIS and some other survey. The implication is that unduplication in a particular block cannot be assured, if one survey is sampling from an area list while another is selecting specific units from a census file.

If unduplication were to be assured some solution to this problem was necessary. Though many alternative solutions were explored, we ultimately settled on converting any unit frame block hit by NHIS into an area frame block to ensure consistency across surveys. (This solution required that NHIS be the first survey to be selected.) The movement of blocks to the area frame was done to the detriment of the other surveys interested in reducing variances through specific sorts of units in the unit frame. (Only block level sorts are possible in the area frame.) Roughly 2 percent to 3 percent of the national samples selected by nonNHIS survey were converted from the unit frame to the area frame because of NHIS "hitting" unit frame blocks.

Because of this, restrictions were placed on how much NHIS sample could be selected in large city blocks.

B. Unit Segments Replace Address Segments

In addition to the unduplication issues, we also decided to replace our address segments with unit segments for all surveys.

Definitions:

Address Segment: Units are clustered geographically. For multiunit addresses, the FR visits the house number and street and lists all the apartments before interviewing.

Unit Segment: Units are clustered according to their characteristics as recorded in the census. Census serial numbers are selected. The specific address, including apartment designation if any, as recorded in the census, is given to the FR to visit and interview. Listing occurs only if the FR cannot find the designated apartment in a multiunit address.

In the current 1980-based surveys, we have an address segment design for CPS, NCVS, and SIPP, a unit segment design for CE, CPP,

AHS-National, and AHS-MS and an all area segment design for NHIS.

So Why Did We Change?

There are two primary reasons:

One reason is to improve our ability to control for duplication of sample units across surveys.

- A unit segment design provides us with a consistent approach for selecting samples across surveys. By selecting specific census serial numbers from the 1990 census, we can control for the duplication of sample units among surveys.

- This duplication control was very difficult to do between address segments and unit segments in the 1980 design. In fact, it was only possible by reselecting sample units for unit segments when they overlapped with sample units in address segments. This has an undersizeable impact on the precision of the estimates that we can obtain for the surveys with unit segments.

A second reason is that unit segments reduce the variances in the survey estimates.

- This reduction is partly due to avoiding duplication through the sample design rather than reselecting sample units.

- The main variance reduction, however, comes from being able to select sample units according to their characteristics as reported in the census.

The success of unit samples in the field depends on how good the list of addresses is in the decennial census files. We intend to review the results of the post-enumeration survey evaluations that deal with the quality of census operations in order to give us perspective on the kinds and amounts of problems to expect.

C. Oversampling

The third major area I want to discuss here is that of oversampling. This is an area where interest is expanding dramatically. We see more and more interest in data on subgroups of the population. This interest has led to increased interest in oversampling to enhance the sample sizes and reliability associated with these subgroups.

Strategies to accomplish oversampling generally fall into two categories:

1. Stratify groups you wish to oversample based on census characteristics. Then sample these groups at a higher rate to increase the likelihood of including cases from these groups.

2. Screen samples at the time of interviews and select cases in the desired subgroup with higher proportions.

The first procedure is less expensive and depending on the level of stratification and the recency of the census data, it can be quite effective. The drawback of course, is that it is difficult to control the level of oversampling and as the time between the stratification and the interviewing grows, the effectiveness of the oversample diminishes.

The big drawback of the second procedure is the cost. It is expensive to screen the cases many of which you do not need data from.

FUTURE DIRECTIONS

As the details of the 1990 sample redesign fall into place we are naturally looking toward 2000 to see what changes we should make. The emphasis of the 1990 redesign was on operations with lighter emphasis on research. In the past 2000 redesign, we see more efforts being directed toward methodological research particularly as it relates to methodological improvements made possible by the automation of the survey process currently unfolding at the Census Bureau.

From the operational side, to major changes will be researched. The first major change that we are working toward is to get to a single frame design. Our current four frames unnecessarily complicate coordination of our surveys. The effects of multiple frames increase the time and money needed to select and modify the surveys. The unduplication effort is probably the most expensive and time-consuming part of the redesign. Moving to a single frame will make unduplication between all list sample survey quite easy. We are also researching the feasibility of seeking relief from the strict unduplication between area frame surveys and list frame surveys.

The emergency of a master address file as part of the 2000 census makes it possible for the need for a new construction frame. If this

can be accomplished it will open the way for a single frame design, which will open the way for a new look at unduplication. The resources saved by these two changes can be redirected toward new and better methodological procedures for household surveys.

SUMMARY

Our experience with the 1990 sample redesign has been very positive. We have coordinated the survey's sample selection and provided a structure whereby we can move into simpler and more flexible redesigns of our survey in the 21st century. This project provides a great example of interdisciplinary work coordinating across disciplines across divisions of the Census Bureau and across several agencies of the Federal Government. Those who worked to make this happen can be justifiably proud.