A MULTIPLE FRAME SURVEY FOR RARE POPULATION ELEMENTS

Joseph Steinberg, Social Security Administration

I. Introduction

This paper discusses one approach to the sample design of a survey for estimating characteristics of a number of rare, but related, target populations. Some of the considerations involved in efficient achievement of study objectives are presented. A number of problems encountered in creating the design are somewhat other than run-of-the-mill and, therefore, may be of interest.

As is described in greater detail both area and list frames are involved in the design. The general plan involved in this sample design is not novel and only in some respects may be considered as having a few new features. In general, when one speaks of using list frames, one envisions certain characteristics--as a minimum, currency of addresses. As a practical matter, many of the lists used for this survey do not have the property of showing current addresses. As a consequence, special procedures are required to insure an unbiased sample. Another characteristic envisioned in the use of list frames is the ability to allocate directly elementary units to first-stage sampling units. As a practical matter, many of the lists (the same ones) can be allocated only to groupings of parts of first-stage units. Again, as a consequence, special procedures are required.

In most plans involving choice between screening for unduplication among frames versus multiple-frame overlap weighting, a decision may be made after consideration of cost and variance. In this sampling plan for the major groupings of list frames and area sample, because of the time relationships of various activities needed for the screening approach, the cost would be extremely large if screening were attempted. Thus, as a practical matter, in these cases only the use of multiple-frame overlap weighting can be used (and at a substantial reduction in cost of securing the necessary information for carrying through the operation). Within one major grouping, screening versus multiple frame overlap weighting exists as a choice; time features would not intrude extensively and as a result after proper consideration of the alternatives screening is being used since it is more efficient.

II. Purpose of Survey

In brief, the purpose of the survey is to study the social and economic consequence of disability among adults. It is estimated that among persons 18 to 64 there are approximately 8% with chronic limitations in their major activities. Among persons in the overall target population, disabilities vary in severity of limitation. The extent of incapacity ranges from limitations in the amount or kind of work which can be performed to total inability for self-care. It is estimated that approximately 1.5% of the adult population is so severely disabled as to be unable to work or keep house. The problems of the disabled population in relation to income insecurity is to be a principal focus of the survey. Thus, the amount of income received by the disabled adult and the sources from which this income is derived (whether major insurance or assistance programs) is of basic interest.

As a consequence of the principal purpose of the survey, six primary target populations were identified. These are characterized by cross-classification of two levels of severity of disability by three categories of source of income. It is speculated that the sizes of these six target populations are approximately as follows:

(In thousands)

Income Source	Unable	Limited
Total	1,500	7,000
Social Security	750	120
Other Public Income Maintenance Programs	600	760
Other	150	6,120

and, therefore, range from about 0.1% to 6% of those 18-64. The primary statistical goals that are to be realized have been formulated as follows:

1. The sample design should permit the determination (at a risk of error of 1 chance in 20) that if 10% of persons in one of the three categories of "unable" have a given characteristic that this is significantly different from an estimated 15% having the same characteristic among another of these three target populations (i.e., that an estimate that 10% of

persons receiving social security disability benefits and classifying themselves as unable to work have a given characteristic is significantly different from a group similarly characterized estimated to be 15% of those who classify themselves as unable to work but who receive no resources from any public income maintenance program).

2. An estimate of any group of approximately 50,000 should be subject to an absolute error of about 10,000 (with a risk of being in error of about 1 in 20).

From this general description, it may be discernible that the usual techniques of area sampling, perhaps supplemented by a single list or two, is not likely to be the most efficient method for dealing with these objectives. A straight area sampling technique would require a very large basic sample (perhaps as much as 250,000 households) to satisfy the study requirements for that target population which is 0.15% of the total population. There are a number of interrelated aspects to the problems of sample design in this situation. Among these are (1) the problems of the availability of certain types of resources such as: Area sample materials and lists. (2) Possible ways of development of special frames. (3) Consideration of double sampling approaches. (4) The possibility of screening among frames for discarding duplicates versus use of multiple overlapping frames with optimal weighting1/ and, (5) Administrative considerations, such as the timing of a variety of activities (i.e., unduplication) as well as the cost factors.

III. The Basic Design

Consideration of the coverage of possible list frames indicated that not all elements of the target populations would be covered by list frames. As a minimum, in order to provide the necessary supplementation for dealing with this gap, it was decided that a multistage, area, probability, double sampling approach would be required. Further, it was decided that the field collection of data in this survey is to be by the Census Bureau. To take maximum advantage of existing sampling materials, the first-stage units are an amalgam of a Census Bureau 197 PSU MLS design and its companion 197 PSU subset of the Census Bureau's CPS--or a 243 PSU first-stage design. Therefore, the 243 PSU's are the areas within which the basic survey is to be done.

Study of the cost factors suggested a ratio of about 20 to 1 between the costs of obtaining and analyzing the detailed characteristics of interest to the field costs of identifying the target populations within the overall population. Pretests have showed the feasibility and moderate costs of a mail plus field follow-up for noninterviews approach as a tool for identifying members of the target populations. Pretests have suggested that among elements of the special list frames 1/2 to 2/3 or more would be members of the target population versus 8% in area sampling. The amount of information to be obtained from respondents, the nature of the other aspects of the survey process (coding and matching to other source data) and the extent of tabulations, suggested a relatively high unit cost for the intensive interview phase of the survey.

The area sampling approach was determined to be sufficiently well accomplished by a first-stage probability design involving 197 strata (243 sample PSU's) with the subsample of units within to be drawn from current survey operations of the Bureau of the Census. (The establishment of the approximate level of the first- and second-stages of the area sampling was part of the joint consideration of components of variability from the area and list frames, as well as costs of all phases such as training, interviewing, matching, etc.) All list samples for field interviewing also have the 243 sample PSU's as the first-stage sampling units. It may be of interest to note that the interaction of availability of area sampling resources and efficiency dictated that in some of the strata (46 out of 197) two probability selections be made. Within the remaining strata 37 consist of a single PSU and in the remainder (114) a single primary sampling unit represents the stratum.

IV. Available Resources and Special Frames

While five of the six target populations are comparatively rare, they vary in difficulty of access.

List frame resources for selection are readily available for those who receive social security benefits (either as disability beneficiaries or as adults whose disability arose prior to their eighteenth birthday). The decisions in regard to sample design between those who classify themselves as unable or

limited are simple for this group. Administration of the Social Security Act requires creation of a number of data files. Thus, tape records showing current residence are available for beneficiaries. Therefore, sampling of the population receiving benefits from the Social Security Administration is quite simple.

For those who are unable and receive some public funds (other than from the SSA) the predominant component is those receiving funds through the Aid to Permanently and Totally Disabled or Aid to the Blind Programs. Effective cooperation by the Welfare Administration and state agencies in sampling the latter frames insures meeting the survey objectives. Localization of effort to lists of current recipients of APTD and AB benefits was deemed the only efficient approach. Other comparatively small groups supplement the APTD-AB to comprise this target population. After some investigation, it soon became apparent that the use of other supplementary special list frames for this population would not be efficient. The sole supplementation for this target population (and for its parallel involving persons with limitations) is through the area sampling approach.

The size of the universe of persons who are limited and without public income maintenance support is substantially large so that the area sample portion of the design, alone, permits meeting the specifications.

Those who classify themselves as unable but are not in receipt of funds from any public income maintenance program are a small and most difficult target population. It is for this latter group that the most extensive sampling effort has been required. Thus, the primary need in the development of special frames arose in dealing with this population. Consideration of likely information sources suggested that some knowledge about this group existed at the SSA. However, it turned out that the information was not readily accessible in a convenient single tape for sampling. In order to deal in some way with this group, three different sources (7 lists) are being used. The information available required (a) sampling of elements within those administrative units containing in part one or more PSU's or parts of PSU's; (b) identification of sampled units to sampled PSU's; and (c) a number of other activities to deal with associated problems. These sources arise in a number of ways. Some persons who become disabled approach the SSA either informally or formally to determine whether they are entitled to benefits under the Social Security Disability Program.

A. Some people who are disabled do not qualify for benefits because of lack of insured status under the Social Security Act (3 lists--for 3 separate years for this group are being used).

B. Others who have insured status do not meet the disability requirements (3 lists--for 3 separate years for this group are being used).

C. Some people make inquiry but do not follow through (1 list is being used for this group).

After some consideration and based in part on pretest information, it seemed possible that perhaps as much as two-thirds of the group of persons unable and without public income maintenance benefits might be covered through these multiple list frames.

In addition to the special sampling efforts mentioned briefly above, use of some of these frames involves a number of special procedures to permit setting up an unbiased design. Special procedures are required since these frames are not current with respect to extent or addresses: They contain members other than those of the target population and do not list current addresses as well as not being directly accessible in terms of basic first-stage units. Nevertheless, the needs of the survey seemed unlikely to be met except through utilization of these frames however difficult the problems might be. (It may be worth noting that the last of these frames could be established only by special intervention in an administrative process so that certain control cards normally discarded were maintained specifically for this survey use. The other frames were accessible through computers but only through the use of data tapes containing skeletonized records.)

The basic identification on each record of these seven special frames is a district office code (and in some cases even this is missing). A district office service area may be part of a county, a county, or several counties. As a result, the first-stage of sampling for these 7 frames entailed selection of

a sample in those district offices which contained any part of a sample PSU. The sampling rate used is the within PSU rate which satisfies the largest requirement for the rate of within PSU sampling.2/ For those elements with no identification, an overall sample was drawn for subsequent subsampling.

Once the sample was selected, addresses were abstracted from case folders and coded to counties. Some units were established as being outside the sample PSU's. These were discarded from the basic design. For the elements coded to sample first-stage units, subsamples were selected to provide essentially a self-weighting sample.

V. Screening Versus Multiple Overlapping Frames

The multiplicity of frames raised the interesting questions as to the desirability and feasibility of screening among frames to permit discarding versus the retention of elements in overlapping domains with optimal weighting. A number of papers have discussed some of the statistical considerations involved.3/ In a sense, this problem is certainly not new. Any use of frames which overlap raises these questions. The multiplicity of the frames involved in this survey and the nature of the information available as well as the cost and time factors involved has led to an approach which combines and uses screening and discarding in some cases, retention of sample elements in overlapping domains in other cases for differential weighting. Among those frames immediately and directly available to the SSA, the screening for overlap and discarding took place by establishing a hierarchal relationship among the nine SSA lists used for this survey. The basic identification of overlap took place by recognizing the sample elements of a lower order frame in the higher order frames through comparisons of social security numbers. The matching process had access not only to the information available on the skeletonized records on computer tape but, where necessary, to the voluminous paper records which support the tape information and are basic to the administration of the Social Security Act.

The overlap among the major groupings of frames (considering all nine social security frames for this purpose as a single frame) can only be established after the fact. The survey process utilizes double sampling for the area sample to identify members of the target population. Immediately after the first phase of the double sampling process, it will be necessary to move to the second, intensive interview phase for all selected target population elements at which time, social security numbers will be part of the identification information secured. Thus, the area sample supplementation of the samples from other frames as well as the interaction between the other frames can only be done after the fact. Social Security numbers will be used for determining overlapping elements between the area sample and virtually all list frames. (Even where there are problems, administrative techniques within SSA for dealing with incorrect or missing social security numbers assure a high degree of success in identification of overlap elements.) For the APTD-AB list frame overlap, data secured in the intensive interview will be the primary basis for identifying the cases in the overlap domains.

VI. Some Other Problems

As is well known, needs for information for sampling are not always identical with information available in a given list frame. Utilization of incomplete or non-current frames involves problems. Frames which cover substantially more than the target population can be dealt with through double sampling. Incompleteness of frames as described above, is dealt with through supplementation through use of area sampling.

Further, in this survey design some complications arose because the skeletonized tape or other records did not show PSU designations but only the overall service area covered by district offices of the SSA. The method for dealing with this problem was through a multi-stage double sampling approach described in Section IV, above.

The major remaining problem encountered in the use of these non-current frames was in the potential bias of failure to deal with the movers. A number of studies, including the Census Bureau's "Reverse Record Checks"4/ have suggested that it is reasonable to expect high levels of success in finding current addresses when starting with older addresses if sufficient field work is done. Based on this premise, we have tackled the lack of current addresses in

these available frames by methods similar to those used in these studies. This problem, the lack of currency of addresses, was approached in several collateral ways. On the one hand, the need was to identify the sample units whose current address is still within a specified sample PSU. Movers, in part, could be sampled by determining the sample elements which had moved to any other sample PSU. For each of these two situations there is then a known probability of selection. The remaining problem lies in dealing with movers who lived in non-sample PSU's at the time of the creation of the record. More specifically, the problem lies in identifying those who lived in non-sample PSU's but who have moved to sample PSU's. To deal with these movers, the approach being followed is to supplement the sample of first-stage non-certainty PSU's with a supplemental sample of PSU's and to select a sample from the same frames in those supplemental PSU's. Those sample elements who are determined to have moved to a specified sample PSU are then part of the basic sample. Thus, the current address determined will, in all cases, be within the original sample of 243 PSU's. To the extent that one can establish current address for those whose addresses were known as of several years ago, we have an unbiased final sample.

VII. Summary

In general, none of the elements of the specified sample design taken by themselves is essentially new. However, some of the indirect approaches for achieving an unbiased design have been deemed to be of more than passing interest. This series of techniques suggests that multiple-frame resource oriented techniques can be found for dealing with rare target populations.

Acknowledgement

The author wishes to thank Harold Grossman for his assistance in the development of the sample design.

FOOTNOTES

1/ Hartley, H.O., "Multiple Frame Surveys" in American Statistical Association, Proceedings of the Social Statistics Section, 1962. Cochran, Robert S., "Multiple Frame Sample Surveys" in American Statistical Association, Proceedings of the Social Statistics Section, 1964.

2/ However, in some cases a lower rate was used to reduce the volume of administrative work involved in the selection of case folders and geographic coding. Where a small PSU with a high within rate was associated in a district office area with a large PSU with a much smaller within rate, a reduced level of sampling was deemed desirable.

3/ Hartley, op. cit.; Cochran, op. cit.

4/ U.S. Bureau of the Census, Evaluation and Research Program of the U.S. Census of Population and Housing, 1960, "Record Check Studies of Population Coverage." Series ER 60, No. 2., Washington, D.C., 1964.