I. INTRODUCTION

The Census Bureau has traditionally redesigned, or at least reselected sample for, its recurring household surveys soon after each decennial census. For the post-1980 census redesign, a major research program is being conducted. Nearly every aspect of the design and estimation procedures is being considered and researched—nothing will automatically be done the way of the past. This does not mean that the ways of the past are inadequate or inappropriate for the 80s, but rather, we are planning the best possible design for our household surveys which will balance survey cost with sample reliability. Much of the design research has been completed, and most of the rest is nearing completion. (Estimation research will continue for about another year.)

A number of decisions have been made, although a number of decisions remain to be made. This paper describes each of the changes in survey design or operation that has been decided upon so far. In each case, the present procedure as well as the new procedure is described, the reasons for the change are briefly given, and quantification of cost or other advantages are given wherever possible. There is no discussion in this paper for cases where it has been decided to make no changes.

The redesign, incorporating these changes, has a number of advantages, even though implementation of the redesign will be expensive. Net savings in survey operations over 10 years in direct costs and/or in the equivalent of variance improvements are expected to exceed $34 million [25]. These savings come about for a number of reasons, but foremost are the changes, including those discussed in this paper, which result from recent research and new technological developments. The savings result not only from efficiency gains but also from being able to meet the primary survey objectives, many of which have changed dramatically in the last several years. Also, in some cases the complexity of the design will be reduced, especially for CPS, making maintenance easier and less error-prone. Implementation of redesign is still uncertain as the required funding is not assured.

Much of the research that has been done has resulted in decisions to continue past procedures. For a description of the full redesign research program, see [7].

Seven surveys are scheduled for redesign: Current Population Survey (CPS), the major labor force survey; Health Interview Survey (HIS), which produces data on health conditions and other health-related characteristics; Annual Housing Survey (AHS), which produces data on components of housing inventory and financial and general characteristics of housing; National Crime Survey (NCS), which estimates victimization rates for a variety of different crimes; Survey of Residential Alterations and repairs (SORAR), which estimates expenditures for residential alterations and repairs for use in the gross national product estimates; Point of Purchase Survey (CPP) which gathers data on where people shop, for use in pricing for the Consumer Price Index; and Consumer Expenditure Survey (CE) which produces data on consumer expenditure patterns. There will also be reserve samples designated for two or three sample designs which are collectively called the General Purpose Survey. This will be available for new surveys that are needed over the next 10 or so years.

One important aspect of the research program has been, for most of the surveys listed above, to designate separate staffs to research each survey on many topics. After the initial research and preliminary conclusions are made by survey, an effort is made to see if common design decisions can be made that are nearly optimal for all surveys. This approach, which has the appearance of being costly, is producing results which far exceed the added personnel and computer costs. In some cases, research done for survey A has even resulted in a better decision for survey B than to which the survey B research by itself would have led. The remainder of this paper describes planned changes. There are separate sections pertaining to the sampling frames, stratification, selection of the sample within PSUs, data collection, estimation and evaluation.

II. CHANGES IN DESIGN, OPERATIONS, AND ESTIMATION

A. Sampling Design

1. Area Sampling

In the current designs, most of the sample for all the surveys is selected directly from 1970 census address listings. In locations where street name and house number are not available from the Census and in locations where permits are not required for new construction, area sampling techniques are used. In general, it has been decided that this basic approach will be continued. A detailed discussion of the advantages of list sampling from the census versus area sampling for Census Bureau surveys will be forthcoming shortly.) However, in the case of the Health Interview Survey (HIS), it has been decided to replace list sampling for the census by area sampling (see section b). Section a. discusses a planned change in one aspect of area sampling that will reduce costs.

a. Allocation of Measures of Size. In area sampling, a clerk divides a sample ED into chunks, with each required to have recognizable geographic features for its boundaries and containing 35 or fewer housing units. A chunk is selected with probability proportionate to size, is listed in the field, and then the appropriate subsample of the listed housing units is selected for interviewing [13].

An expensive part of preparing an area segment for interviewing is allocating the units enumerated in the census to the right chunk of a sample. This allocation is used as an aid in assigning a measure of size to each chunk, where the measure of size is based on the proportion of units allocated to the chunk. In preparing for the 1978 Registration and Voting Survey, a procedure was developed that requires substantially less clerical time for allocation. It consisted of using those housing units spotted on the map by the census enumerator and allocating the unspotted housing units equally to the land chunks. Although the ultimate area segments may be more variable in size, this method cuts the time...
required for allocation almost in half. It has been tentatively decided that this new methodology should be used for the redesigned surveys. The net savings for this methodology over the old, after allowing for the effects on variance, are estimated to be greater than $1.2 million over a 10-year period.

b. Health Interview Survey (HIS). For the HIS, there are some special considerations which favor area sampling over list sampling. The sponsor, National Center for Health Statistics (NCHS), would like to use expired HIS sample for other surveys they plan to conduct during the decade. Based upon current procedures which select sample addresses directly from the 1980 census listings, private companies would be prevented for confidentiality reasons from conducting these surveys. If area type segments were listed specifically for the HIS, they would be exempt from Census Bureau confidentiality rules, and follow-on surveys could be conducted by the organization most appropriate for the activity. NCHS has concluded that the potential savings and advantages of using the HIS sample addresses for some of their other surveys outweigh the disadvantages. Thus, we plan to use area sampling rather than census list sampling as the main source of sample for HIS. (See [9] for more details.)

2. New Construction. A new construction unit is one that has been built since the decennial census from which the sample addresses are selected, i.e., for the redesign, samples this would be since April 1980. Most of the sample of new construction units is taken from a "permit new construction" frame. The permit new construction frame consists of permits issued by the Census Bureau, and 2) smaller offices which report an average of two types for sampling purposes -- 1) those that are larger and more active, which generally make monthly reports to the Census Bureau, and 2) smaller offices which report annually and for sampling purposes are presently assumed to issue permits at an average of two units per month. Sampling intervals are applied to the cumulated permits to yield sample permit offices. The addresses of all permits issued during the month are listed at the permit office. Sample addresses are selected from permit address listings and sent out for interview as permit segments [13].

a. Unavailability of Permits. The new construction sampling for the redesigned survey is scheduled to begin in 1983. Thus the initial redesign new construction universe will consist of building permits issued over the span of about 3½ to 4 years. Between the time the permits are issued and the permit addresses are listed by the Census Bureau, the permit information may become unavailable for use (i.e., the permits may be lost or destroyed or they may be filed in such a way that for a specific permit month they cannot be readily identified). The unavailability of building permit information can cause problems in the redesign in several ways. We are doing something new in the redesign implementation to alleviate the most serious of these, namely, the effects on the initial redesign sample selection. The other types of problems are discussed in [1].

If steps are not taken to modify the new construction sampling procedures in areas where the permit information may be unavailable, the population living in newly-constructed housing units in those areas will have no chance of selection. The partial solution decided upon for this problem is to predetermine "problem" permit offices, and to then collect data from these offices while it is still available rather than wait until 1983. This was not done in conjunction with the 1970 redesign. At that time whenever there was a hit in a particular permit office and permit month for which building permit data were not available, a "similar" sample permit was selected to be double-weighted. This increased the variance slightly, and to the extent that the "similar" segments were different from the segments for which they substitute, a bias was introduced. The predetermination of "problem" offices should greatly reduce the frequency of double weighting, thus reducing both bias and variance. For more details, see [1].

b. Sample Selection of Permit Offices. After considerable research and thought, it has been decided to make some significant changes in the methods for handling the sampling of permit offices. The major change is for the smaller offices. Analysis has found that the average number of permits issued per month is under 2, and that there is great variability in the number of permits an office may issue within a given year. Under the present system, a lot of time and money is spent on assignments in which interviewers are sent to permit offices only to find that no permits were issued that month. The change will be that actual annual reports of permit activity will be used instead of assuming activity of two units per month for sample selection purposes. Sampling will be done only once a year with assignments being made to list a full year's accumulation of permits rather than just two month's. This procedural change will thus reduce the number of assignments made for listing permits, although it will also result in a slight loss of data quality since more time can elapse between issuance of the permit and sampling. The estimated net monetary savings over 10 years in 1981 dollars is $130,000. For more details, see [24].

c. New Construction Shortly Before April 1980. Units for which new construction permits were taken out after April 1980 will of course be included in the permit new construction frame. Further, units for which permits were taken out several years before the census should certainly have been enumerated in the census and thus need not be included in the permit new construction frame. There is a problem, however, with units for which permits were taken out a few months prior to April 1980. Some of these units were ready for occupancy in April 1980 and thus should not be included in the permit frame since they were in the 1980 census; some were not yet finished and, thus, can be sampled only by inclusion in the permit frame.

For the present sample design where 1970 census sampling materials are used, a uniform cutoff date of January 1, 1970 was used. It was found during the 70's, however, that there were about 600,000 units which had no chance of selection
under this system. Thus, a special sample of these missed units was selected at extra expense and introduced into the surveys in the mid and late 70's [12].

A research program was instituted to avoid this problem after the 1980 census and to determine an optimum starting point for the redesign permit sampling. The primary goal was to have approximately equal numbers of units which have two chances of selection (i.e., included in both the 1980 census frame and the permit frame) and which have no chance of selection (included in neither the 1980 census nor the permit frames). A secondary goal was to keep both of these quantities as small as possible.

It was found that there were substantial differences depending on structure size and region of the country and thus starting points are to vary by size of structure/region cell. The optimum months for starting the sampling of monthly reporting permit offices, annually reporting permit offices, and public housing, are given in [15]. Details on the research methodology are given in [12] and [15].

The change to these new starting points should result in smaller bias for CPS as compared to what was done in the 70's, since the state supplements to CPS were not adequately covered by the special sample of missed units. A final decision to use starting points was made rather than use of a later starting point in conjunction with selection of a special sample of missed units, as was done in the previous redesign, has not been made pending comparisons of bias and cost.

d. Clustering of Sample Permits. Currently, permit addresses are assigned map grid coordinates and these coordinates are then used to geographically cluster the addresses prior to selecting clusters of permit addresses for sample. The clustering is done so that units assigned for interview will be as close together as possible, thereby reducing the amount of travel. A study has been undertaken to investigate this procedure and to determine in particular whether it would be more effective to assign enumeration district (ED) numbers rather than grid coordinates to the addresses and, if so, whether it would be better for the permit address seeker or the computer to make the assignment. (EDs are administrative areas used in the 1980 census which average about 300 housing units in size.) A tentative recommendation has been made to assign permit new construction to EDs and to have this done by permit address seeker. [7] [16] The final decision will be made after further study of some additional cost data for alternative clustering procedures [17], and subject to availability of funds for map acquisition.

Assignment of ED numbers has several advantages over map grid coordinates. The full list of advantages is given in an internal memorandum [16]. As an example, one of the advantages will be mentioned here: ED boundaries are stable throughout the decade, thus eliminating problems that now occur in preparing interviewer materials when an area adopts a new set of maps resulting in grid coordinate changes.

e. Computerization of Permit Sampling. At the present time, the entire permit sampling operation is done clerically. A study has been made to determine whether all or some phases of the operation should be computerized. There are three phases of the operation: (1) creation of the sampling frame, (2) the sampling of permit months and the permits themselves, and (3) designation of sample units and preparation of the field interview materials. It has been decided that all three of the phases should be computerized for monthly reporting permit offices. The major reason for computerization is cost savings in clerical time--it is estimated that savings of over $1.6 million in 1981 dollars will occur during the decade for all the surveys as a result of computerization. A second major reason is that much tighter control of the sample should be possible, making the system flow smoother and be more responsive. See [18] for more details.

3. Representation of Census Misses and Inadequately Described Addresses. Addresses in areas sampled from census address listings that were missed or inadequately described in the census have been represented in the current designs for most surveys by a special supplemental sample. This sample was identified through a process in which an independent sample was intensively canvassed and then matched to the 1970 census [13]. This is an extremely high per unit cost of sampling materials if a special sample is conducted just for this purpose. However, some of this is usually done to evaluate census coverage. A study has been done for several major surveys to determine the impact of this special sample on survey estimates. In each case it was determined that the impact was negligible because the proportion of the U.S. population covered by this sample is only about 1 percent [19], [20], [21]. The percentage of units missed in the 1980 census is believed to have been even smaller than the percentage missed in the 1970 census, so that the impact on survey results would be expected to remain small after redesign. Thus, it has been decided to do no special canvassing for the purpose of representing census misses and inadequately described addresses in the redesigned surveys. However, units of these types are being identified as part of the census evaluation, and will be used as much as possible in the redesigned surveys. It has not yet been decided how the units identified in the evaluation will be allocated among the surveys.

B. Stratification

The stratification of PSUs in the current design was originally done in the early 1950's. The stratification has been periodically revised and expanded since then to reflect changes in population size and distribution and changes in other characteristics among areas. (See [13].) Complete reordering of strata definitions leading to new independently selected sample PSUs has been avoided in previous redesign efforts so that the number of new interviewers to be hired and the number to be fired could be kept small. The strata were formed for purposes of national and regional data, particularly for use in the Current Population Survey (CPS), although they were constructed using a wide range of economic and demographic data so that stratification could be relatively efficient over a wide range of subject matter.

1. Within-State Strata for CPS. In the mid-1970s, data needs for CPS changed considerably;
estimates at the state and substate level become very important. Strata which cross state boundaries and which were formed for purposes of producing national estimates are somewhat inefficient for producing state estimates. There is also a widespread, though erroneous, perception that unbiased estimates are possible only from strata and categories where the stratification was specifically done within states. Thus, to improve efficiency and perceptions, the redesigned CPS will be designed and stratified on a state-by-state basis, with the goal of meeting state reliability requirements at the lowest possible cost while also being sufficient to meet national reliability requirements.

2. Stratification Methodology. The decision to form state-based strata in CPS necessarily means total revamping of existing strata, at least for CPS. Thus, research has been conducted on the best methodology for stratification. Consideration was given to "traditional" (rectangular) stratification and stratification by cluster analysis. For several reasons discussed in [6], it was decided that rectangular methods would not work very well in this situation.

Cluster analysis is a collection of techniques for exploratory data analysis. It works with sets of objects and tries to find natural grouping among them. Beginning in the late 1960s, several different cluster analysis algorithms were developed.

After considerable research, we decided to use a modified version of an algorithm developed by Friedman and Rubin [3]. (The algorithm we will use is very similar to that developed by Dammstrom and Hagnell [2].) The modified algorithm iteratively reallocates PSUs until any single PSU move will not result in a further reduction in between PSU variance. Details on how we decided what methodology to use are given in [6].

Plans are to use this algorithm separately for each redesigned survey. Depending on the results, there may be commonality of strata definitions between surveys, although this is not a pre-set condition. Thus, if significant gains can be made by using different strata definitions for different surveys, it will be done.

Sample size reductions of as much as 1/3 in some states would be possible for CPS due to use of the improved methodology and to our ability to form strata specifically with state data in mind. The equivalent monetary savings is about $9.7 million over 10 years. Additional monetary savings of over $2.5 million are expected from the use of the new methodology for the other redesigned surveys.

C. Selection of Sample Within PSUs. The first step in selecting a sample within a PSU is to designate census enumeration districts (EDs) from which clusters of units will be selected. In the present design, EDs are sorted into four categories: C, central city; B, non SMSA; R, the urbanized area not in the category C; U, urban place, not an urbanized area and not in category C; and R, all other EDs. Within each of these categories, the EDs are sorted in a way which tends to place geographically contiguous EDs together in the sort. EDs are then sampled systematically with probability proportionate to their measure of size. Finally, a cluster of units is selected from each sample ED. Cluster size averages four units for most surveys. See [13] for more details.

1. Sorting-Enumeration Districts. Considerable research has been done to determine the best method for sorting EDs. A variety of different straight sort procedures have been tested for each of the major surveys, as well as the use of the modified Friedman-Rubin clustering algorithm used for stratification. It was not predetermined that each survey would use the same sort procedure and variables—if no one procedure worked well for all surveys, then different procedures would be used for each survey. It was found, however, that one procedure and choice of variables was optimum or nearly optimum for all surveys.

The tentative plan is to form clusters of EDs using the modified Friedman-Rubin clustering algorithm. The methodology also rank-orders the EDs within a cluster as well as the clusters themselves. Different variables will also be used in the clustering depending on the population of the PSU. In the research using the 1970 census data, the population groupings were over 500,000, 125,000 to 500,000, and less than 125,000 people. Slightly different groupings may be used for the actual ED sorting using 1980 census data. For information on the research methodology and more detail on the results, see [8].

2. Cluster Size. Final decision for cluster sizes have not yet been made. It is likely that for some surveys the cluster size will be smaller than at present and for some time. However, interviewers have been required to attempt to conduct first, second, and fifth month interviews for a unit. (Fifth month has been done personally because there are 8 months between the fourth and fifth interviews for a unit.) Beginning December 1981, interviewer instructions were changed for the second month interviews to encourage
telephone rather than personal interviews. This change was made apart from redesign to reduce operating costs. However, the decision to make this change was based on experience in the Methods Development Survey. This is a survey research project for testing alternative collection methodologies and concepts for CPS, and is indirectly part of the redesign research program. In this survey, second month interviews were conducted by telephone without apparent problems or effects on survey estimates [11].

E. Estimation--Current Population Survey (CPS) Composite Estimate. There will likely be a number of changes in estimation and weighting for all the redesigned surveys, but research is generally in the early stages in this area. We do have one almost final decision, however, regarding CPS. At the present time, a simple composite estimate is used for all labor force data for monthly estimates and for averages over time. The basic idea of the composite is to form a weighted average of two different estimates. One estimate is the simple direct one based only on the current month's data. (All stages of adjustment and ratio estimation are included, however, in this estimate.) The second estimate takes the preceding month's composite estimate and adds to it the estimate of change based on the 3/4 of the sample which is in common between the preceding month and the current month.

Although a final decision hasn't been made, we expect to change to a more complex form of the composite estimate for monthly labor force data. (The estimator to use for annual average state data is still uncertain.) This new estimator is referred to as the AK composite estimator. The basic difference between this estimator as it would be used and the present composite estimator is that an additional factor is used in the estimator which has the effect of assigning a greater weight to the part of the sample which is not common between the current month and the preceding month, and a smaller weight to the part of the sample which is common. (See [4] for the exact form of the estimator.) Reference [5] has detailed comparisons between the two composite estimators. Gains of about 2½ percent in variance and, depending on the assumptions one is prepared to make, significant gains in bias, are possible. The equivalent of over $2,000,000 in cost savings are expected over 10 years due to the adoption of the AK composite estimator.

F. Evaluation--Reinterview Program. A portion of the sample in most Census Bureau surveys is reinterviewed. In reinterview, we attempt to uncover dishonest errors made by interviewers, to uncover honest errors, to estimate response variance and response bias, and to deter interviewers from shortcutting interviewing procedures. At the present time, an interviewer is generally scheduled to interview portion of his/her work reinterviewed a fixed number of times in some time interval. For example, up until July 1982, a CPS interviewer was scheduled for interview 1 month in the first half of a calendar year and 1 month in the second half. She/he had no way of knowing which month reinterview would take place, but reinterview would never take place more than once in a 6-month period. Thus, it was possible (though probably rarely happened) for an interviewer to take advantage of this in making dishonest errors, e.g., fabricating all or parts of interviews, in the months following that in which reinterview took place.

To alleviate this problem, the reinterview has been made less predictable. For example, for CPS the new strategy will assure that an interviewer is selected for reinterview a minimum of once per year. However, an interviewer may be selected several additional times. The selections beyond the first will be done at random, so that different interviewers will be reinterviewed different numbers of times across the years, with there being no fixed pattern within a given year. (There will be a maximum of four selections of an interviewer per year.) It is expected that this new strategy will make reinterview more effective in deterring and detecting dishonest errors. It will have little effect on the fulfillment of the other reinterview objectives. See [23] for a detailed discussion of this new strategy and how the decision was reached to implement it.

III. SUMMARY

Section II described a number of changes that have been decided upon for the redesign of the Census Bureau's recurring household surveys as the result of recent research. Significant gains in cost, variance, bias and ease of field operations will result from the planned changes. The overall net gain in operating costs or the equivalent in improvement in variance over 10 years, after accounting for the cost of implementing the redesign, has been conservatively estimated at about $34 million in 1981 dollars [25]. This estimate includes some gains from changes that are anticipated but have not yet been decided upon and thus are not discussed in this paper.

IV. ACKNOWLEDGEMENTS

Extensive improvements were made in this paper as the result of thorough reviews and comments by about 20 people, mostly from within the Census Bureau. These changes described in the paper are the result of the research efforts of a large group of people in Statistical Methods Division and other parts of the Census Bureau. The author, however, is responsible for any errors in this paper. Also, special thanks to Edith Oechslin for typing several drafts of this paper.

* * * * * * * *

REFERENCES


REFERENCES (continued)


[15] " " " " 2/1/82 " " Permit Lag Committee.


[17] " " " " 9/1/81 " " E. Ann Vacca.

[18] " " " " 2/18/81 " " Ruth Ann Killion


[21] " " " " 11/2/81 " " James Dorsch.

[22] " " " " 6/4/82 " " Elizabeth T. Huang.

[23] " " " " 3/11/82 " " Irwin Schreiner
