PLANS FOR 1980 CENSUS COVERAGE EVALUATION

David V. Bateman and Charles D. Cowan
Bureau of the Census

I. INTRODUCTION

The purpose of this paper is to present the objectives and methodology of the 1980 census coverage evaluation program. The emphasis is on issues pertaining to the estimation of the census undercount of persons. In particular, two methods for obtaining these data will be discussed: demographic analysis and a post-censal sample survey, the Post-Enumeration Survey (PES).

Several research projects pertaining to the methodology to be used for the post-censal sample survey were conducted as part of the census pretest in Oakland, California, and in the dress rehearsal censuses conducted in Richmond, Virginia and southwest Colorado. These research projects are briefly discussed. In addition, the possible use of matching with "independent" administrative records in improving the census undercount estimates was tested by utilizing the February 1978 Current Population Survey (CPS) and the Richmond census data. These studies are also discussed.

Since a demand now exists for census undercount estimates for relatively small geographic areas, a section of this paper will discuss regression-synthetic estimation techniques that could be employed in providing these estimates.

In addition, an evaluation of special census coverage improvement procedures will be discussed.

II. OBJECTIVES OF THE PROGRAM TO ESTIMATE THE CENSUS UNDERCOUNT

The two primary objectives of the 1980 Census Coverage Evaluation Program are:

. To develop estimates of the coverage of population and housing in the census.

. To evaluate specific census operations as to their effect on census coverage.

A. Estimates of Census Coverage

1. Demographic Analysis - The demographic method (demographic analysis) of census evaluation involves developing expected values for the population at the census date by the adjustment and combination of demographic data from sources essentially independent of the census being evaluated and comparing these expected values with the census counts. The particular method that is used for demographic subgroups depends on the nature of the available data. For ages under 45, in 1980, estimates will be developed on the basis of birth, death and immigration statistics. For ages over 65 aggregated medicare data will provide the basis for estimates of coverage. For the remaining age groups an analysis of all censuses since 1880, along with death and immigration statistics, provides the basis for developing coverage estimates in 1980 [1]. Demographic analysis will provide national estimates of net census errors for age, sex and race groups. These estimates are measures of net error for age, sex, and race groups, combining coverage errors and errors of content. The demographic method is considered by census staff to be more effective than a post-census sample survey for developing satisfactory estimates of net census errors at the national level for the total U.S. population. However, problems do exist with demographic analysis; the major one is the estimation of the number of undocumented aliens. At the present time, no definitive methodology is available for including this segment of the population in the demographic estimates.

Demographic analysis will also provide state estimates of net census errors for broad age categories, by sex, and for white and black racial groups. However, it is questionable whether they will be better estimates than those produced from the post enumeration survey, and to what extent they will be utilized.

2. Post-Census Enumeration Survey (PES) - The data does not currently exist for using demographic analysis techniques to provide reliable estimates of coverage error for subnational geographic areas such as cities, SMSA's and revenue sharing areas; in addition, the data now available for demographic analysis cannot provide estimates of coverage error for some important socio-economic categories. The Census Bureau will conduct a sample survey as soon as possible after the census enumeration has been completed in order to fill this void. Persons listed in the PES are matched on a one-to-one basis with the census listing of names. Census resources exist for conducting a PES that would provide reliable estimates of net coverage error at the state level for the total population, and at broader geographic levels such as regions or divisions for race-ethnic origin categories. Furthermore, the PES will enable methodology to be developed (e.g., regression-synthetic estimation techniques) that might provide reasonably accurate estimates of coverage error for demographic, socio-economic categories at the state level and for the total population at sub-state area levels (large cities, SMSA's, some revenue sharing areas, etc.).

Post Census Enumeration Surveys were conducted as part of the 1950 and 1960 census evaluation programs. However, the results of these studies were considered not to be successful for providing accurate estimates of the undercount for certain subgroups of the population. It was determined on the basis of other evidence (births and death registrations, plausibility of the obtained sex ratios, observations on the sources of underenumeration and on age misreporting trends) that the PES estimates of underenumeration were seriously biased downward. This was especially evident for black males ages 15 to 59 where the PES yielded a net undercount estimate approximately one-half the estimate provided by births, deaths, migration data, and previous censuses. One can conclude from these results that persons enumerated in the census are much easier to enumerate in the PES than persons missed in the census; that is, persons missed in the census will not be reported in the PES for the same reasons that they were missed in the census. This problem is often referred to as "correlation bias."

The emphasis in conducting the 1950 and 1960 post enumeration surveys was on obtaining PES data of good quality. Highly qualified staff were hired, given extensive training, and a considerable amount of time was devoted to seeing that procedures were properly conducted. The effect was to reduce errors due to poor enumerators and carelessly implemented procedures; however, the correlation biases arising from the tendencies of certain segments of the population not to be enumerated were largely unaffected (in fact, they may have been increased).

The emphasis in the 1980 PES will be on independence from the census, in addition to quality. In addition, the 1980 PES will utilize "independent" administrative files for purposes of improving the estimates of coverage error. To the extent that a satisfactory match between the administrative files and the census and PES can be achieved without impairing independence of the sample data, we should be able to obtain more accurate estimates of coverage error than were obtained in 1950 and 1960. Two administrative files are being considered: the IRS tax return file for persons aged 17 to 64 years of age and the Medicare file for persons 65 or over. The feasibility of using these administrative files is being investigated in a study currently underway. Data were collected from the persons in the February 1978 Current Population Survey in order to facilitate a match with administrative files; the Current Population Survey (CPS) is being used as a proxy for a nationwide PES. Dual system estimates of the true total population will be made as of February 1978 and compared with estimates based on births, deaths, and previous censuses. If the two estimates of total population are reasonably close and the processing problems of administrative file matching are surmountable, administrative files will be used along with the vital statistics estimates, to adjust the PES estimates of coverage error in the 1980 census.

a. PES Methodology

Since the post enumeration survey will begin after the census has been completed (to maintain independence of the two operations, and to insure that the census is conducted properly), different procedures can result from trying to associate census day residence(s) for persons that come into the PES sample.

Three different procedures for doing the 1980 PES have been considered. All three procedures involve an independent canvass of housing units in a sample of areas, a listing of persons in sample housing units, and a search of census records to see if they were enumerated. The methods differ with respect to the rules for linking persons to sample housing units and the form of the estimate used. The listing of persons can take the following form:

- Procedure A (PESA) - A listing is made of all persons who resided at the sample housing unit at the time of the census. These census records for the sample address are then searched to see if the persons were enumerated.

- Procedure B (PESB) - A listing is made of all persons currently residing in the sample housing units together with all persons who died in these households subsequent to the census. A determination is made where each listed person was living at the time of the census. These addresses are then searched in census records to see if the sample persons were enumerated.

- Procedure C (PESC) - A listing is made both of persons who resided at that housing unit at the time of the census and persons currently residing in that housing unit. A determination is made for each current resident where they were living or staying at the time of the census. Thus both PESA and PESB information is obtained from persons residing within the selected housing units. However, searching to census records is only done for persons living there at the time of the census.

The 1950 and 1960 post-enumeration surveys used procedure A. This procedure has both problems of recall (as to census day occupants) and a potentially large noninterview rate due to persons who have moved out of the selected residence since the census.

Since the PESB procedure is only concerned with obtaining a roster of persons at the current address, we would expect this procedure to yield a more complete listing and a better estimate of undercoverage than was feasible under Procedure A, but at a higher cost due to additional matching costs.

Procedure C represents an attempt to combine the best features of Procedure A and Procedure B. If research indicates that the matching rates for PESA nonmigrants is similar to the matching rates for PESB nonmigrants, (intuition would indicate that they should be) a "reasonably accurate" matching rate is obtained for Procedure A movers, while Procedure B obtains a better estimate of number of movers, then Procedure C could be a viable alternative in 1980. This procedure involves conducting the matching operation with Procedure A information separately for migrants and nonmigrants and applying the resulting match rates to migrants and nonmigrants estimates obtained from Procedure B to obtain a coverage error estimate; this would avoid the costly matching and geographic coding operations that are associated with Procedure B.

One of the key elements for Procedure C is a "reasonably" accurate estimate of the matching rate for Procedure A outmovers. A significant bias on this estimate could result from failure to report outmovers on a selective basis. Consideration is being given to the possibility of matching on a subsample of PES B movers as a contingency plan in case a "good" matching rate on Procedure A outmovers is not obtainable.

In addition to estimating a gross omission rate, we also plan to estimate erroneous enumerations in the census; therefore, the purpose of the PES will be to estimate a net coverage error, gross omissions minus erroneous enumerations. For both Procedure A and Procedure B two alternative definitions exist for defining "missed" and "erroneously enumerated" persons.

. Definition I - A person is "correctly enumerated" if he should have been enumerated and was enumerated once and only once, even though it might have been in an incorrect location. A person is "missed" if he should have been enumerated in the census but was not enumerated in any location. An enumeration is considered to be an "erroneous enumeration" if the person should not have been enumerated but was (e.g., he did not exist, lived outside the U.S., was born after the census or died before the census), or the person should have been enumerated but was enumerated more than once.

. Definition II - A person is "correctly enumerated" if he was enumerated in the census at the address reported by the PES as the census date residence. A person is "missed" if he was not enumerated at the census date residence that was reported in the PES. An enumeration is considered to be "erroneous" if the PES reports that the person was not living at the location where the census recorded him. For example, the PES could report that no such person exists, or that the person was born after the census, died before the census or was living elsewhere on census date.

b. PES Research

Post Enumeration Surveys were conducted as part of the census pretest in Oakland, California and the census dress rehearsals in Richmond, Virginia and southwest Colorado. The major reason for conducting these post enumeration surveys was to develop and test appropriate methods for estimating the 1980 census undercount. Thus, individual operations were analyzed closely, resulting in a considerable delay in producing undercount estimates. The major issues to be resolved from the pretest and dress rehearsal post enumeration survey results are outlined below.

1. Oakland Census Pretest

The post enumeration survey conducted as part of the Oakland census pretest was selected from the census mailout list of addresses and units identified as missed in the Housing coverage check. The following issues are being evaluated:

. Procedure A vs. Procedure B - Separate samples, with appropriately designed questionnaires, were selected in order to test Procedure A against Procedure B. No attempt was made to obtain a PES interview from PESA outmovers who did not leave a forwarding address with the sample households present occupants and the present occupants knew little about the census day occupants. Thus an effective evaluation of Procedure C cannot be done.

. Area Sample vs. List Sample - Indications are that the list sample approach (sample selected from the census lists) has sufficient difficulties from an "operational" and an efficiency standpoint to warrant an area sample in 1980.

. Followup to obtain additional matching information - The feasibility of a followup survey in 1980 on a subsample of match and nonmatch PES cases is being evaluated in Oakland. Since all nonmatch cases were followed up in the Oakland PES, we will be able to determine the number of cases that were converted to match status. (A followup was not done for match cases as the matching rules were sufficiently strict to restrict erroneous matches to an absolute minimum, and no attempt was made to estimate erroneous census enumerations.) This research will enable us to develop optimum matching rules for 1980.

. Questionnaire Design - Experience with the Oakland and Richmond processing operation indicates that a great deal of matching difficulties could be alleviated if the matching were to be done on the PES form. We are now designing and testing such a form.

2. Richmond and Southwest Colorado PES's

. Sample Design - A sample of "blocks" was listed and enumerated for the PES's conducted in Richmond and Southwest Colorado. Design effects at the block level will be calculated as input to the 1980 PES sample design.

. Procedure A vs. Procedure B vs. Procedure C - Information on current residents (for Procedure B) and outmovers was collected from the households in the PES sample. While the outmover data were collected primarily for use in the check for overenumeration, the data will permit evaluation of Procedure C as an alternative method of estimation in 1980. If one assumes the equivalence of Procedure A and Procedure B for nonmovers, one can also make a Procedure A estimate.

- Definition I vs. Definition II for Procedure B - Sufficient information was collected and is being tabulated to evaluate this issue.

- Followup of movers - For definition II, one of the major problems with checking for overenumeration is the difficulty of locating enumerated outmovers to obtain information on the residence at the time of the census that is comparable to that for missed inmovers. In addition, for the Procedures A and C samples (to estimate omissions), socio-economic and administrative record information will be collected; a substantial nonresponse rate from proxy responses due to movers could result in large biases of undercount rates at the subgroup level. A search for Procedure A outmovers was done approximately 11 months after the census. We are now in the process of evaluating the results of this study.

- Classification problems - The Oakland, Richmond, and Colorado surveys enabled the Bureau to acquire a considerable amount of experience in the handling of PES noninterviews, PES and census records with insufficient matching information, and census imputations.

3. Multiplicity Research -

Multiplicity or network sampling was tested extensively in the Oakland, Richmond and Colorado surveys. In a regular household survey, an individual can be reported only by one household, usually the one of which she/he is a current member. If some or all persons in the population can be reported by more than one household, the survey would constitute a network sample. Thus, as in the Richmond-Oakland PES, Procedure C will usually call for a network sample, where the counting rule would include all current resident members of the sample household plus any other persons who had lived there at the time of the census. In addition, as part of the post enumeration surveys conducted in the above areas, network surveys with extensive counting rules were tested to determine their feasibility for estimating the census undercount. Sample household members not only reported their current de jure household members, but also reported specified relatives who live at different addresses. These relatives were then matched to census records to see if they were enumerated. A followup of reported relatives was conducted in order to determine if underreporting of relatives was a serious problem. Counting rules for siblings, parents, and children were tested and are now being analyzed. Preliminary indications are that, except for children, we were unable to obtain sufficient address information (for matching purposes) for a significant number of cases. Furthermore the inability to obtain good matching information appears to be concentrated in minority groups, the exact area where improved methods for estimating the undercount is needed. At the present time, we are considering the possibility of including a multiplicity counting rule on the PES questionnaire for a limited subrule (if such a subrule exists) that will provide good matching information and for which underreporting has been, historically, a major problem.

B. Evaluation of Specific Census Operations as to their Effect on Census Coverage

A number of special procedures and operations are being designed to improve the coverage of the 1980 census. Some of these procedures and operations have been used in previous censuses; however, a large number are to be used for the first time in 1980. Since these programs have a significant cost associated with them, and they occupy a great deal of staff time in their implementation, it is felt that they should be evaluated from a cost-benefit and a quality standpoint. Specifically, the evaluation would consist of measuring the improvement in coverage resulting from the operation, relative to its cost, and a determination if the operation was correctly implemented. Relevant data on the operations will be collected in specified district offices. The operations and procedures that have been targeted for an evaluation are:

1. Precanvass: The precanvass is a dependent field check in which an enumerator annotates a copy of the purchased tape address register for an enumeration district (ED) by canvassing that ED, making additions, deletions, and revisions in the address register as necessary.

2. Nonhousehold Sources Program: The purpose of this program is to utilize lists of names and addresses from sources not associated with the census, and then match these lists to the census. After matching a followup is made for persons from the source, who are not found in the census, in order to determine if the person was missed in the census. If missed, the person is added to the census. As a further coverage improvement procedure, the entire census day household roster is recorded at the time of followup and a check is made to see if they were enumerated in the census. The sources to be evaluated are driver's license, and U.S. Immigration and Naturalization Service (INS) Alien Registration files.

3. Misclassified Occupied Units Study: All units classified as vacant or deleted by the census will be followed up so that any true occupied units can be identified and the occupants enumerated.

4. Census Questionnaire Coverage Items (Q1 and H4): These questions are designed to identify potential missed persons and housing units.

5. Dependent Roster Check: Certain households are being followed up because their questionnaires failed certain coverage and content edits. A census day roster will be obtained from these households and a comparison will be made with the roster on the census questionnaire. Persons missing from the questionnaire will be added to the census.

6. Whole Household Usual Home Elsewhere: A circle on the questionnaire is filled by the respondent if the whole household has a usual residence elsewhere (URE). If so, a procedure is established to add the household to the census in that URE area. The original housing unit is determined to be vacant, and the household is identified as temporary residents for that area.

7. "Were You Counted?" Program: Forms are provided for the public through newspapers and other means to enable persons, who feel they were not enumerated in the census, to be enumerated. For forms that are sent in census household rosters are checked to see if they were enumerated.

8. Casual Count Program: An independent list of names and addresses is made in locations where persons with a high probability of being missed in the census might congregate (e.g., bars, pool halls, etc.). These names and addresses are matched to census records to see if they were enumerated.

9. District Office Local Review: A printout of preliminary population and housing counts is generated after the census second followup stage. "Major" differences between 1970 data and 1980 census counts of population and housing are flagged. The differences are adjudicated in the field so that appropriate actions can be taken.

10. Local Officials Local Review: Housing and population counts for local areas at the block, ED, tract, and jurisdictional level will be submitted to local officials for review while the district offices are open. Arrangements will be made for a proper investigation and a correction process if required.

11. Tract Block Deletes: These type of deletes are addresses that occur within a given area on the commercial mailing list but are not within the Tape Address Register (TAR) boundaries for that area. These addresses are deleted since it is assumed that the prelist operation will pick them up. Errors can occur if they are not deleted and prelist picks them up, or if they are deleted and prelist fails to pick them up.

12. Post Enumeration Post Office Check (PEPOC): Address listing for conventional areas are submitted to the post office for a review of completeness. Followups (and resulting additions to the census counts) are made when appropriate.

13. Effectiveness of Assistance Centers: Centers are being developed to provide assistance to persons unable to read or fill in the census forms.

III. MATCHING TECHNIQUES

One of the most difficult operations to design and implement is the development of matching techniques that involve:

- matching of PES housing unit and person records to census enumerated housing units and persons.

- Matching of PES and census enumerated housing unit and person records to "administrative" file records.

These matching operations are different in that the former involves a searching operation in a file arranged by address, whereas the latter involves searching files arranged on some other basis (in the case of the IRS and Medicare files the search is on the basis of a social security number). Therefore, our research effort has taken different paths in determining optimum procedures for these two operations.

A. Matching of PES Housing Unit and Person Records to Census Records

The matching operations conducted for the Oakland, Richmond and Colorado post enumeration surveys were clerical in nature with explicitly written matching rules. The Oakland PES was our first attempt to create a set of matching rules; since they were changed a number of times during the experiment, a definitive set of rules does not exist for Oakland. Based on our Oakland experience a set of explicit rules for persons was devised for Richmond and Colorado. The basic matching operation consisted of the following:

1. Coding the PES addresses to tract, ED, block, serial number, and form type. This information is needed to locate the address register and the corresponding census questionnaires.

2. Matching PES listed housing units against the census address register in order to obtain an estimate of census housing unit coverage. Maps with corresponding map spotted units were used when searching for census addresses. Also the block header record that identifies the ED and block for a given street name and house number proved to be very useful when searching for census addresses. Telephone and city directories were used to a lesser extent in the searching operation.

3. Transcribing information from the PES interview forms to a special form to be used to control and facilitate the person matching. (Note we are considering dropping this operation for the 1980 PES, and doing the matching directly on the PES questionnaire.)

4. Matching persons on the match forms to persons on the census questionnaires. Name, relationship, sex, age, date of birth, and race were used as matching variables for Richmond and Colorado.

5. For the Oakland PES, all Procedure B nonmatches, and "possible" match cases were followed up to see if additional information could be obtained to determine match status for the "possible" match cases or to obtain additional

address information for the nonmatch cases.

6. Lastly, a final matching operation to census questionnaires was conducted to determine final match status.

The following are general observations based upon our experience with the matching operations:

. The matching operations will require some form of validation in 1980. This could include all possible matches and a subsample of match and nonmatch cases.

. Followup (or reconciliation) will involve only cases for which additional PES information is needed to determine match status. If the additional information cannot be obtained, the case will be included as part of a noninterview adjustment and a search for a corresponding census record will not occur.

. Matching Procedure B inmovers has been a difficult task. Indications are that we were unable to locate a significant number of reported census day addresses (addresses other than the PES address); also, many addresses that were located were done so only with a great deal of difficulty.

We are now investigating the possible use of computer matching. Due to the impossibility of keying the entire census, and the lack of a FOSDIC name, this would probably involve keying census records to those blocks or ED's that have PES sample. Thus, clerical matching would still be used (because of census geocoding errors, movers addresses outside the sample blocks, etc.) but to a much lesser extent. This might prove to be especially feasible in rural areas where good address information is often lacking. Our major concern with this method of matching is timing. Setting up a major keying operation could delay the completion of the PES processing up to three months beyond what a 100% clerical operation would take.

B. Matching of PES and Census Enumerated Housing Unit and Person Records to "Administrative" File Records

Certain groups of persons are particularly likely to be missed by both the PES and the census; examples are: black males, males in urban "ghetto" areas, low income adult males and migrants. Two administrative files are being used to provide alternative estimates to the PES-Census match coverage estimates for these groups. These files are the Internal Revenue Service (IRS) tax return file for persons of ages 18 to 64 and the Medicare file for persons of ages over 64.

The methodology to be used in forming a "triple system" census coverage estimate will consist of matching PES records and a sample from census enumerations to the IRS and Medicare files. Triple system estimation is explained later in this presentation. Matching will be done on the basis of a reported social security number. The Social Security Administration's alphadat and Summary Earnings Record File will be used to obtain social security numbers for census and certain PES records, and to validate reported PES numbers. This is discussed more fully in Section IV.

IV. ADMINISTRATIVE MATCHING

A possible improvement to using the PES to estimate net undercoverage in the census by a match to census records (dual system estimation) is to additionally match to administrative records to form triple system estimates. The two sources planned for use in 1980 are the tax returns filed in 1980 for 1979 fiscal year and the Medicare file of all Medicare records for the year 1980. There are several problems with using these files. The IRS tax file alone contains about 85 million records, stored on 131 data tapes in social security number order.

Names and addresses are given to the Bureau exactly as they are listed on the tax return, meaning the address could be the address of the tax filer's bank, lawyer, or whoever prepares his tax return. The Medicare file is similar, but on a smaller scale. Thus information may be reduced for confirming or negating matches.

To match to either of these files, it is necessary to have a social security number for the record to be matched. Note that this is true for records matched to the IRS or Medicare files, but not necessary if matching is done from either file to the census. The distinction will be clearer in a moment. The reason for needing the social security number is twofold:

1) Since the files are in order by social security number, it is most cost effective to search thre files using that indicator. Matching to these files using names or other variables would be prohibitively expensive.

2) The social security number is nearly a unique identifier. While one person may have several social security numbers, possessing more than one SSN is a relatively rare event, and on the IRS files each SSN should belong to only one individual. However, identification using a person's name and matching in either direction can have problems when the individual possesses a common name (e.g., Robert Smith).

Unfortunately, for these purposes, social security number is not collected in the census, even on a sample basis. To match census records into the social security system, either an SSN has to be obtained for census records by matching census records into the Social Security Administration's name file (a Soundex system), or one can take records already matched to the census which have the SSN available. The PES sample (which includes the nonmatch cases) can be used for this purpose, as it is already a state sample (for total corrected population). To obtain the full triple system estimate, cases that were discovered in the census that were not enumerated in the PES must be added to the PES sample and matched to administrative records.
Matching can go in the other direction, too. A

sample of cases with name and address can be drawn from the IRS and Medicare files and matched back to the census, in much the same way the PES is matched to the census. However, problems with matching in this direction arise due to the need for a timely state sample; special arrangements would have to be made with IRS to draw a state sample while they are receiving return forms. This is necessary because the final IRS tax return file with names and addresses isn't available to the Census Bureau until approximately a year after the receipt of the forms.

It is also anticipated that a followup operation would be necessary because of the portion of the sample from IRS which would list an address used for tax return purposes which was not the residence as of Census day. This could introduce a substantial bias into the dual system or triple system estimates by causing a low matching rate at the person's residence.

Research is being conducted now to determine which direction is less problematical. A supplement was administered as part of the February 1978 CPS, collecting information necessary to matching the sample into the IRS tax return file for fiscal 1977. Dual system estimates will be developed from this matching project which will be compared to demographic estimates for 1978. This project should give us an indication both of the problems to be encountered in matching in this direction and will also tell us, by comparing the dual system estimate to demographic estimates, whether the assumption of independence of sources in the dual system estimate holds.

The other project being attempted is a match of 2700 sample cases from the IRS tax file to the census records for the Richmond, Va. and Colorado 1978 Dress Rehearsals.

This project will indicate what problems there are matching from IRS to a census, and will also be compared to the PES match results to see what differences there are in the two population estimates. If possible, a similar project will be attempted with the Medicare files before 1980. However, we anticipate relatively few problems with this match (as compared with an IRS match) in 1980.

V. ESTIMATION

The primary emphasis of the estimation procedure is to provide estimates of the net undercount for states (including the District of Columbia). A primary goal of the coverage evaluation program is to provide a methodology for determining corrected population counts at the state and substate area level. Since we cannot afford a survey to accomplish this objective at the local area level, we are developing a program that could be utilized in developing synthetic regression estimates at this level. Broadly speaking, this will involve a PES sample that will provide reliable estimates of the corrected population of specified minorities for broader than state area levels (e.g., region, district, or urban-rural level). The first estimates that could be formed after the census is concluded, would be dual system estimates of the total corrected population for each state and for certain large SMSA's and cities. To obtain these estimates, the Post Enumeration Survey will be matched back to the census, with the match status ascertained for each person in the household. The PES sample is being drawn as a state sample with supplementation of the largest SMSA's and cities. Within each state or SMSA, the last stage of sample selection will be blocks or a subsample within blocks to facilitate the matching process. Each person or household in the sample will ultimately be classified as correctly enumerated, omitted, or erroneously enumerated. The sample estimates of the proportion of matches and of erroneous enumerations will be used in the dual system estimate to obtain the total corrected population in each of the states and designated SMSA's and cities.

The dual system estimate is basically that used in capture-recapture methodology to provide population counts of migratory animals, birds, and fish. Of necessity, one or two modifications have been introduced to allow for the vagaries of survey data. The estimate is formed as shown in Table 1, below.

Table 1: Forming a Dual System Estimate for One of the 61 Divisions

		Census		
		In	Out	Total
Post Enumeration Survey	IN	M'	-	N'_P
	OUT	-	-	-
	TOTAL	N'_C	-	N'_T

where $N'_T = \dfrac{N'_P \cdot N'_C}{M'}$ is the dual system estimate of the total corrected population for one of the 61 divisions.

N'_P is the estimate from the PES of the total population, uncorrected;

M' is the estimate from the PES of the number of persons enumerated in both the PES and the census;

N'_C is the total population count obtained in the census, minus the estimate of erroneous enumerations and of the total number of imputations made.

The only assumption required in this model is that the two sources be independent. If independence holds, then N_T is the maximum likelihood estimate; N'_T is the final estimate of the total corrected population. It already allows for processing errors, census refusals and other cases which could not be matched since the cases are represented in N'_P but not in M'. To estimate the completeness of the census count or to estimate the census undercoverage, we must add the imputations and erroneous enumerations back to N_C. That is

$P_c = \frac{N_c}{N_T'}$ = estimated completeness of census enumeration

where $N_c = N_c' + E_c' + I_C$ = actual census count including erroneous enumeration (E_c) and imputations (I_C)

also $w_c = \frac{N_c'}{N_T'} = \frac{M'}{N_P'}$ = proportion matched estimated completeness of the actual field enumeration, excluding erroneous enumerations and before any imputations.

Imputations and erroneous enumerations have to be excluded in estimating N_T because none of the imputations or erroneous enumerations will be matched and thus will not be included in M'.

Also using the above notation

$O_c' = N_T' - N_c'$ is the number of persons not counted in the census ;

$O_c = N_T' - N_c = O_c' - E_c' - I_c$ is the difference between the total corrected population and the census count ;

$q_c = 1-p_c = \frac{O_c}{N_T'}$ is the net undercoverage rate ;

and $r_c = 1-w_c = \frac{O_c'}{N_T'}$ is the gross undercoverage rate.

These procedures can be found in Marks, Seltzer and Krotki [2]. Following the work of Deming and Chandrasekaran [3], the dual system estimate is formed for demographic subgroups within the region for which the estimate is being formed. These estimates are made for the smallest mutually exclusive demographic categories (e.g., young black males), and added across categories to obtain the estimate for the region. This is done to reduce both the variance and the bias of the estimate.

These estimates would be revised as more information about the undercount becomes available from administrative record matching. Matching will be done using administrative records, and separate estimates of the undercount can be formed from a Census-IRS match and from a Census-Medicare match. These would be compared to the Census-PES estimate and an adjusted estimate prepared. Demographic estimates for the U.S. as a whole will also be available. The state estimates obtained from matching can be adjusted to these national totals. As mentioned previously, there are timing problems in obtaining estimates from matching to administrative records, which lead to these estimates being produced later than the PES estimates; hence the need for revisions.

A more complex estimator can be formed which involves a good deal more work. The concept of the dual system estimate can be expanded to comprise an n-system estimate, where now three sources are used in the matching process: the census, PES, and a combination of Medicare records and the IRS tax return file. Matching problems faced in the dual system estimate increase threefold because of the number of relations possible. Offsetting the increased matching problems, however, gains are made in both reduced variance and reduced bias when employing three systems. This is illustrated in work done by Woltman and Smith [4] and Wittes [5]. The final step in the estimation procedure is to produce estimates at lower levels of geography. The sample for PES is being designed to produce reliable estimates of the total corrected population for each state and the ten largest cities and SMSA's. But there is an interest in producing estimates at the county level, and possibly at the revenue sharing area level. Producing estimates at the county level is more probable than producing estimates at the revenue sharing area, because there are only 3,300 of the former, but over 41,000 of the latter, most of which are very small. To produce these estimates, two alternative methods are being considered: regression estimation and synthetic estimation.

Research is now being conducted to compare the advantages and disadvantages of a regression vs synthetic approach. For regression estimation for counties in which we have sampled, estimates will be formed of the net undercount for each county. These will be used in conjunction with demographic data collected in the PES and census to develop models of the net undercount. Research into this area is looking at what variables are important to the model, what alternative regression models might be used, and what transformations on the data might be needed. For the synthetic estimates, alternative synthetic techniques are being compared as well as the level of aggregation to which the estimates are being made (Purcell [6], Gonzalez and Hoza [7]).

VI. SUMMARY

A large scale sample survey will be conducted as soon as possible after the 1980 Census with the purpose of estimating census population and housing unit counts, corrected for the undercount. The survey is presently being designed from the standpoint of appropriate methodology, including an optimal sampling plan.

Results from the post enumeration surveys conducted as part of the censuses of Oakland, Richmond, and southwest Colorado are presently being analyzed. The results of these studies will determine, to a large extent, the methodology to be used in the 1980 PES.

Administrative records are being considered as a part of the estimation process. Studies involving the February, 1978 Current Population

Survey and the Richmond census are being conducted to determine if and how they can be used.

Results from the Oakland, Richmond, and southwest Colorado census - PES match studies are being analyzed to determine optimum matching rules.

Techniques are being developed for estimating corrected population counts for subnational area levels, and for small demographic - socioeconomic subgroups in large areas.

BIBLIOGRAPHY

[1] U.S. Bureau of Census. Census of Population and Housing: 1970. Evaluation and Research Program. PHC(E)-4, Estimates of Coverage of Population by Sex, Race, and Age: Demographic Analysis, 1973.

[2] Marks, Eli S., William Seltzer, and Karol J. Krotki. Population Growth Estimation. A Handbook of Vital Statistics Measurement. New York: The Population Council, 1974.

[3] Chandrasekaran, C., and W.E. Deming. On a Method of Estimating Birth and Death Rates and the Extent of Registration. Journal American Statistical Association, 1949, pp. 101-115.

[4] Woltman, Henry and William Smith. An Internal Census Bureau Memorandum, Preliminary Finding on Dual vs. Triple System Estimation. June 4, 1979.

[5] Wittes, Janet T. Application of a Multi-nomial Capture-Recapture Model to Epidemicological Data. Journal American Statistical Association, 1974, pp. 93-97.

[6] Purcell, Noel J. Efficient Estimation for Small Domains: A Categorical Data Analysis Approach. Unpublished PhD Dissertation Biostatistics Dept., University of Michigan.

[7] Gonzalez, Maria Elena and Christine Hoza. Small Area Estimation with Application to Unemployment and Housing Estimates. Journal American Statistical Association, 1978, pp. 7-15.