The 1990 Post-Enumeration Survey: An Overview

Howard Hogan, Bureau of the Census, Washington DC 20233*

Key Words: Undercount, Census, Matching, Dual System Estimation

1. INTRODUCTION

The 1990 Post-Enumeration Survey (PES) will constitute the major vehicle for measuring coverage differentials by area in the 1990 Decennial Census. If the results are of sufficient accuracy, the PES will form the basis for any adjustment of the Census that might occur. This paper presents the statistical concepts underlying the PES design. It addresses the sample design, the interviewing instrument, the matching concepts, the missing data imputation methodology and the method of smoothing the results to reduce sampling variance. It presents the method, and the justification for the chosen procedure.

2. SAMPLING

The PES is a sample, in this case a sample of 170,000 housing units in approximately 5,400 sample block clusters. A block cluster is either one block or a collection of several small blocks. To be useful to correct the Census enumeration, the PES results must be generalized to people living in the non-sample blocks. To do this, the population is divided into groups, or post-strata. The Census count is known for each group. The PES estimates the true population for each of these groups. Thus one can calculate the ratio of the PES estimate of the true population to the Census count. This ratio is called the adjustment factor. The post-strata are based on the following characteristics:

Race: Black, Non-Black Hispanic, Asian and Pacific Islanders, and all other
Age 0-9, 10-19, 20-29, 30-44, 45-64, 65+
Sex: Male, Female
Census Division: New England, Mid-Atlantic etc.
Place/Size: *Central City of Major Primary Metropolitan Statistical Areas (PMSA's)
*Central City of Large Metropolitan Statistical Areas (MSA's) (with at least one city with population of 250,000 or more)
*Central City of Small MSA
*MSA: Not Central City
*Non-MSA incorporated places with population of 10,000 or more
*All other
Tenure: Owner vs. Renter

Tenure is included because previous research has indicated that renters, especially urban minority renters, are especially difficult to count. (Isaki et. al. 1987).

Given the objective of producing estimates of the population for the post-strata, the sampling strata should correspond to the post-strata (defined by all variables except age and sex) as closely as possible.

The cross-classification of divisions and the place/size categories above yields 54 major geographic areas that will serve as major sampling strata. The next step involves creating additional sampling strata within these areas by grouping geographic units with high concentrations of the race-origin-tenure groups corresponding to the post-strata for that geographic area. For this purpose, 1980 Census counts of occupied housing units by tenure and the race-origin of the householder were used. Finally, a sampling stratum was created having a large proporation of American Indians. This stratum is defined to include persons living on Indian reservations. After the grouping of geographic units, 101 sampling strata were defined.

A few groups have been excluded from the PES frame. People living in institutions have been excluded, as have military personnel living in barracks and people living in remote rural Alaska. Finally, the survey excludes the population defined by the Street/Shelter operation, "S-night." [See Alberti et. al. (1988) and Woltman and Alberti (1988) for a complete description of the sampling.]

3. LISTING AND INTERVIEWING

Maps were prepared for each of the sample blocks. These maps included none of the information gathered for the Census enumerations. PES field work began in February when "Current Survey" (i.e. non-Decennial) interviewers visited each of the sample blocks to list all housing units that they contained. This listing was done out of Regional Offices by permanent Census Bureau field staff.

PES interviewing started in late June 1990. The Census non-response follow-up was scheduled to finish by early June. Other Census operations were conducted concurrently with the PES. Interviewing was largely conducted by former enumerators. However, to help ensure independence, the PES was managed out of twelve regional centers rather than out of the local District Offices.

During the summer, interviewers visited each housing unit in the sample blocks to find out who is living there and where they lived on April 1, 1990, Census Day. There are two aspects of the interview that were given special attention.

The first element is coverage. The PES interview must not exclude the very people missed by the Census enumeration. To the extent that the PES systematically misses the same groups of people that the Census misses, it will underestimate the Census undercount. This is the bias due to response correlation.

The interviewers ask about the people living there at the time of the PES interview. They do not try to

reconstruct the household as of Census Day. We decided that it is hard enough to interview the "hard-to-count" where they are living at the time of the interview. It is unreasonable to expect to interview them months after they have left.

The second concern is non-response in the initial PES interview. Some people will not be at home. Others will refuse. If the people who refuse the Census also tend to refuse the PES, downward correlation bias will be introduced. Some people who answered the Census may now refuse, thinking they have helped enough, which will cause a bias in the opposite direction. In any case, failure to obtain a good interview will introduce uncertainty into the PES estimates.

The PES began three months after Census Day. Many people have moved. To determine whether they were enumerated in the Census, we must know where they lived on Census Day, as well as where they live now. We try to get this information in two ways.

First, the interviewer asks:

How long has ... lived or stayed here?

If the person has lived there more than a year, the interviewers goes on to the next person. If he has lived there less than a year, the interviewer asks:

What date did ... move to this address?

If the answer is before April 1, the interviewer goes on. If the person cannot quite remember, the interviewer probes:

Did ... move to this address before April 1, 1990?

Then finally the interviewer asks:

What was ...'s address on April 1, 1990?

There is a weakness with this approach. People who move in and out, as well as people with multiple addresses, sometimes misinterpret the questions. For example, if you ask a mother when her son moved into her home she might well reply "He's always lived here," even though he was away at college in April. So everyone is asked:

People sometime have more than one place where they stay. This can cause us to count them more than once. Did any of the people now living here stay any part of March or April of this year
at a college or university
with another relative
at a second home
on a military base or ship
somewhere else for any reason?

In spite of the questions and probes, some people who moved in after Census Day will say that they did not. Assuming that they were counted at their Census Day address and not at their current (sample) address, they will be falsely classified as not-enumerated. This error will lead to an upward bias in the measured undercount.

4. MATCHING

The goal of the PES matching is to produce the correct ratio of cases classified as omitted to those classified as included in the Census. To accomplish these goals the PES processing is guided by several concepts.

First, the matching classifies people as included in the Census only if they were counted at the address where they should have been counted, according to the information they provide. We call this concept "Correct Address Matching" or "Unique Address Matching." For example, Census rules require that a college student be enumerated at the university dormitory, not at his parents' home. The PES will count the student as "enumerated" only if he is counted at the university. If he is not counted at the university, then the student is classified as "omitted" even if he was counted at home. In order for the estimation to work out, we must classify the enumeration at home as erroneous and subtract it from the Census. So in this example, we would have one omission (at the university) and one erroneous enumeration (at home). The two net out in the aggregate.

The second concept is that of the search area. If a person reports that he lived at a given address, then the matching classifies him as correctly enumerated if he was counted anywhere in the block. It will also classify him as correctly enumerated if he was counted in a surrounding block. We do not want to classify a person as missed if he was counted next door or across the street. However, there is a limit to how far the matching process can search. If a Census operation has coded the address across town, for example NW vs. SE, the matching will not search there and will not find the person. The matching will count him as missed. To balance, the system must count the other enumeration as erroneous, because it is outside the defined search area.

A final concept is the idea of "Sufficient Information for Matching." When a match is found, it is easy to say that the case was enumerated (although not necessarily *correctly* enumerated.) When no match is found, it may not prove that the person was not enumerated, but merely that we are not looking in the correct place. A further review of the case might show that we have "insufficient information," leading to its being classified as "Unresolved." There is a strong tendency to classify cases that match as "Resolved: Enumerated" and cases that do not match as "Unresolved." This can create a strong bias. Because of this, the rules that classify cases as "Sufficient Information for Matching" are applied before the matching begins. These rules are designed so that if we find a match we will be confident that the person was correctly enumerated and, equally important, if we do not find a match, we will be confident that the person was omitted. This leads to a somewhat higher "unresolved" rate, but also to more accurate overall results.

The first stage in matching is done by the computer. This is a complex process that we have developed over the decade. See Jaro (1989), Winkler and Thibaudeau (1990) for a complete description of the theory and implementation. Computer matching began as soon as both the Census and the PES files

are virtually complete for that District Office.

A match may exist even though the computer matching system does not fnd one. Clerks and technicians review all possible matches and unmatched cases. Clerks can take account of relationships, can review notes, and can decipher handwriting. The computer matching system prints out the results by household. If any individuals match, all other Census household members identified in the Census are printed in one column and all other household members identified in the PES are printed in another. All cases undergo 100% independent quality control.

The clerks first check the sample block for matches. They then search each of the surrounding blocks. To aid them, alphabetized lists of all people in surrounding blocks are provided, together with the actual Census questionnaires.

In fact, the clerks do not just classify the cases into Unresolved and then into Enumerated or Omitted. We have an elaborate classification system. The Omitted cases are classified into:

Within Household Misses
Housing unit included but whole household missed
Whole Structure Misses
Census Processing Error (i.e. questionnaire returned but not counted in the Census.)

Some cases will not undergo computer matching. These are the Census cases that were data captured after PES computer matching began. These will be sent directly to the clerks for clerical matching.

The accuracy and consistency of the matching process are central to the PES process. The rules have been developed over a decade of research. However, the task of controlling such a large operation conducted in such a short time frame is one of the major challenges ahead.

Full names are not data captured in the Census except for those people enumerated in the PES blocks and surrounding blocks. The computer matching works only for people who were living in the sample or surrounding blocks on Census Day. For people who moved since April 1, we must rely on another process. Clerks must assign the reported Census address to a Census block. If this cannot be done with some confidence, clerks classify the case as "Insufficient information for matching" and send it to follow-up. If the addresses can be coded, clerks must access the microfilm image of each housing unit in that block. If the clerk still does not find the person, then they normally stop and make sure that we have not geocoded the cases incorrectly. This may require follow-up.

An important part of the treatment of movers is confirming that the search is being conducted in the correct area. For example, the respondent may report correctly that he was living at "1102 Elm," but the interviewer may record, "1012 Elm." Even if the interviewer records the response correctly, there may be an Elm Street, an Elm Avenue, an Elm Court and an Elm Terrace. Before the matching process classifies a person as not-enumerated, clerks must confirm that they are searching in the correct area. To aid in this process, the PES interviewers have asked several questions about the reported April 1 address:

What are the names of the cross streets, roads, highways, or other landmarks closest to that address?
What are the names of two neighbors living near that address?
Was there anyone who lived there [at the alternative address] *on April 1 who does not live at this address now? What are their names?*

If the clerks find the cross streets, locate the neighbors, or find any of the other household members, then we know that we are looking in the area the respondent reported. If they do not find any of this confirming information, then they must either try to re-code the address or send the case out to get more information.

5. MEASURING ERRONEOUS ENUMERATIONS

The process described so far measures Census omissions. However, omissions are only part of coverage errors. To measure net undercoverage, one must measure both gross omissions and gross erroneous inclusions. Thus, the PES actually consists of two samples. The Population or P sample measures the proportion of people included in the Census. The Enumeration or E sample measures the proportion of enumerations that are correct. Erroneous inclusions include Census duplicates, Census fictitious enumerations, people who were born after Census Day or who died before Census Day, and people who were counted in the wrong place. The E sample consists of all Census enumerations coded, correctly or incorrectly, to the blocks sampled for the P sample. For purposes of *sampling,* it does not matter where the person, housing unit or address actually was, only where the Census coded it.

The design treats an enumeration as correct if, according to the information provided, the person should have been counted either in the sample block or in one of the surrounding blocks that make up the search area. The process takes several steps:

First, clerks must search throughout the block and the surrounding blocks to see if the case was a duplicate of another enumeration. If a duplicate is found, one of the enumerations is erroneous. Second, they must see whether the person actually lived at the address on April 1 and whether the address was actually in the search area.

For every enumeration in the E sample that is linked to a person interviewed during the summer, interviewers have already asked the questions about where that person lived on Census Day. For every address in the E sample that is also in the P sample, interviewers have already "spotted" it on a map. The enumeration may not be classified as "correct" just because the pair matches. A person may have already reported that he moved in after Census Day or was away at college. Actually, to reduce E-sample follow-

up, during the June PES interview, interviewers have already asked whether there was anyone who lived at the address on Census Day who did not live there now. These people are not in the P sample. However, if they were enumerated, then they can be classified as correctly enumerated.

There will be people enumerated in the Census who were not interviewed by the PES interview in the summer. They may not have been at home. Their house may have been missed by the PES. They may simply have been excluded from the PES roster. They may have left and not been reported by the current resident. Interviewers must ask them the same questions that were asked in the P sample: where were they living on April 1, 1990? Were they away at college, at a second home, etc.? Then we apply the same rules that we apply to the P-sample people to determine the "correct" April 1 usual place of residence. If the building where they were living was not picked up in the PES listing, then interviewers must see whether it is in the PES sample block or a surrounding block. The interviewer must "spot" the true location on a map. If it is outside the search area, clerks will code the enumeration as erroneous.

However, there is an important class of Census enumerations where no one expects to reinterview the people in the PES. These are fictitious Census enumerations, "curbstones" in the jargon of the Census Bureau. Proving that someone does not exist is not easy. When the interviewer asks

"Do you know Dottie..?"

an answer of "No" may indicate no more than that respondent does not know Dottie, not that Dottie does not exist. The rules require the interviewer to find at least three knowledgeable respondents in an effort to determine whether an enumeration was fictitious.

One final type of enumeration bears special mention. The Census sometimes includes enumerations with such sparse data that it is impossible to determine the unique individual referred to. An example is enumerations without names. Other examples are cases where the Census only determines that a unit was occupied but not who was living there. Even if the people were included in the P sample, it would be impossible to match accurately to these enumerations. Thus, for PES matching and estimation, all these cases are classified as "Insufficient Information for Matching" and treated as not in the Census. Of course, these cases are included in the Census counts when computing net coverage error or applying the adjustment factors.

In processing the E Sample, it is important to include all Census enumerations, especially those conducted long after April 1. We have a special operation to process Census enumerations that were data-captured after computer matching. This operation presents special challenges in merging the data with the results of the earlier operation and completing the processing in time for follow-up. However, it presents no new conceptual problems.

6. FOLLOW-UP

As mentioned above, sample Census enumerations that were not linked to a P-sample case are sent to the field for interviewing. In addition, certain P-sample cases that have not been matched will be sent to the field for follow-up. These include the following cases:

Whole household non-matches with conflicting information: cases where the Census reports one family (the "Emersons") as having lived at the address on April 1, but the PES interview reported another family (the "Petersons") as having lived there on Census Day. Experience from the 1986 PES test has shown that sending the non-matched P-sample cases out for reinterview together with the E-sample people leads to more accurate reporting of both. (See Hogan and Wolter, 1988).

Whole household non-matches without conflicting information: cases where the PES interview has indicated that the family lives at the address, but the Census enumeration lists the house as "Vacant" or includes it as a Census non-response.

Whole housing unit non-matches: cases where the PES interview has indicated that the family lives at the address, but the Census enumeration has omitted the unit or coded it outside the search area. In these cases, it may be that the unit was vacant but the PES people misreported the fact that they moved in after April 1.

Non-matched Proxy Interviews: cases where the initial PES interview was with a neighbor or other non-household member.

Non-Matched Movers: The problems with identifying and geocoding the actual April 1 address can lead to the clerks looking in the wrong location and thus failing to find a match. In some cases we will have confirming information that shows we are in the correct location. However, the operational problems of sorting these out before Follow-up has lead to the decision to send all non-matched movers to Follow-up.

It will be important to maintain a low non-response rate during Follow-up. Because only difficult cases are sent to follow-up, they are far more likely to have been missed in the Census than cases selected at random. Failure of Follow-up could result in losing the very cases we are most interested in including in the PES.

We expect Post-follow-up processing to begin as soon as the first cases return from Follow-up and continue for about four weeks after the end of Follow-up. It will include a final P-sample match, including mover matching, as well as coding the E-sample cases as correctly or erroneously enumerated.

After matching, the final "processing" step is merging the data records created by the matching clerks with the initial data files for both the P sample and E sample. One might think that this would be a simple process, but with approximately 400,000 records to merge on both sides, problems and delays should not be a surprise.

7. ESTIMATION

The first step of non-response adjustment is weighting responses for the whole household non-

interviews. Then, there are two stages of item non-response adjustment.

One stage consists of imputing for missing demographic characteristics and other variables that will be used for post-stratification (see below). For example, if race is missing, we will impute it based on the race of those members of the household, or the neighbors. If age is missing, we impute it based on the distribution of the response cases.

The more critical phase of non-response adjustments corrects for missing enumeration status: correctly enumerated vs omitted for the P sample, correctly enumerated vs erroneously included for the E sample. This is done by a logistic regression fit to the response categories. The carrier variables include:

Age group
Sex
Race
Hispanic
Whether the case went to Follow-up
Tenure
Mover/Non-mover
Proxy
Before Follow-up whole-household non-match
Before Follow-up within-household non-match

Separate P- and E- sample regressions are run. For the P-samples, mover status is also a carrier variable.

For post-stratification, we are now able to include:

Age and sex

We can now define race/Hispanic and Tenure based on the PES report: Multiplying all the cross-classification variables together would give almost 2,000 post strata. Some of these are so small that we collapse them into "adjacent" cells. For example, there is no tenure category for rural areas, nor separate category for Blacks living in rural New England. After all collapsing, we retain 1,392 separate post-strata.

The Dual System model conceptualizes each person as either in or not in the Census enumeration, as well as either in or not in the PES.

		ENUMERATION	
PES	Total	In	Out
Total	$N**$	$N*_1$	$N*_2$
In	N_1*	N_{11}	N_{12}
Out	N_2*	N_{21}	N_{22}

All cells are conceptually observable except for N_{22}, and of course any of the marginal totals that include N_{22}. However, if the PES is an (approximately) unbiased sample of the whole population, then an (approximately) unbiased estimate of $N**$ can be made.

$$N** = (N*_1 * N_1*) / N_{11}$$

This is the so-called dual system or Petersen estimator of total population.

The post-stratification serves two purposes. First, we would like to know the undercount or overcount rate for each of the groups. This is important for estimating the net undercount at the local level. Also, both theory and previous experience indicate that the model is a better approximation if the estimate is computed for more nearly homogeneous groups. (Sekar and Deming, 1949; Wolter 1986).

It is acceptable for both the PES and the Census to have different coverage rates for different post-strata. However, if within a post-stratum, there are sub groups where both the PES and the Census have significantly lower coverage, then the dual system estimate will be biased. This is one type of correlation bias.

Another type of correlation bias would arise if being enumerated affected the person's response to the PES, or being in the PES affected the person's response to the Census enumeration. This would be the case if the PES interviewer and the enumerator compared notes, or if a person refused to cooperate in the Census Follow-up because he had been recently interviewed in PES. Conducting the PES after most of the Census operations are complete and conducting the PES out of the Regional Census Centers rather than out of the local District Offices that conduct the enumeration should minimize this effect.

Note that $N*_1$ is the number of distinct and identifiable people in the Census. This differs from the official Census count which includes duplicates, fictitious cases, and other erroneous inclusions. These are measured by the E sample and subtracted before forming the estimates. The difference between the estimated true population, $N**$, and the Census count (including now erroneous enumerations) estimates the net Census undercount. The ratio of the estimated true population and the Census count is called the adjustment factor.

8. SMOOTHING AND COMPARING

The PES sample size of 170,000 housing units may seem large, but it may still suffer from unacceptably large variances for subgroups of the population. On average, there are over 250 people per cell, but in fact the smaller cells may have many fewer. To control for the variance, we have instituted a smoothing program.

Essentially the Census Bureau will fit a model to the raw adjustment factors. The final carrier variables will be determined by the data, but they will be limited to data that appear on the Census 100 percent data file. The model will include indicator variables such as age, sex, race, and tenure. It may also include the number of whole person substitutions required in the Census processing, which could serve as an index of the difficulty of enumeration. The model is used to predict the adjustment factor for each of the post-strata.

After the model is fit, the predicted factor is averaged with the raw factor with the weights inversely proportional to the sampling variance and the model variances.

The Census Bureau is considering ways to combine this PES estimate with estimates from auxiliary data sources, chiefly from demographic analysis. For example, demographic analysis estimates of the ratio of Black males to Black females by age at the national level may be superior to those given by the PES. The Bureau is

investigating methods to use this auxiliary information to improve the PES by ranking the individual post-strata estimates to agree with demographic analyses at the national level for those groups where the demographic analysis estimates are superior (Bell and Diffendal, 1990). We would then recompute the adjustment factor.

9. ESTIMATING AT THE BLOCK LEVEL

The adjustment factors are the ratio of the estimated true population for an age, sex, race, etc., post-stratum to the Census count for that group. To correct the Census then requires multiplying the factor by the Census count for any area. This is done at the block level. We choose the block to insure that all tabulations based on the adjustment are consistent.

The adjustment will generally not produce whole numbers of people to be added. Neither the Census tabulation and publication system nor the majority of Census users is prepared to deal with fractions of people. We will round fractions either up to a whole person or down to no person, using a controlled procedure. That ensures that the total for the block is not rounded up more than one or down more than one.

The PES post-strata employ broad age categories, 0-9, 10-19, etc. It employs only four race/origin categories: Black, non-Black Hispanic, Asian and Pacific Islanders, and all others. To adjust the Census, we want to add detailed records. For example, if we are to impute a 0-9 year old Hispanic in a predominantly Mexican-American origin block, then the process will impute an exact age (say, 5) and, usually, impute the person to be Mexican-American. If there is no one of a given ethnic origin, age group or sex, in a block, the PES cannot add anyone of that origin, age or sex there. This imputation will be done using a "Hot-Deck" procedure similar to that already used in the Census.

10. CONCLUSION

The 1990 Post-Enumeration Survey involves many stages of sampling, interviewing, matching and estimation. Each stage must have several steps, and may undergo quality control. Many people have labored to design this survey over many years. It reflects the Bureau's experience dating back to 1950. More recently, it reflects the lessons learned in the 1980 Post-Enumeration Program and in the 1985, 1986, 1987, and 1988 test Post-Enumeration Surveys. We have endeavored to build in quality and accuracy, aiming to produce adjustment factors by May 1991 (and final adjusted tabulations by July 15, 1991, if necessary). Although the survey has already begun, much of the effort lies ahead of us. Only time and further analysis will tell whether we have succeeded.

References

Alberti, Nicholas et. al, 1988, Preliminary Stratification Schemes for the 1990 Census Coverage Measurement Programs. Paper presented at the Joint Advisory Committee Meeting, October 13-14, 1988.

Bell, William and Diffendal, Gregg, (1990), Proceedings of the Section on Survey Research Methodology of the American Statistical Association.

Hogan, Howard and Wolter, Kirk (1988), "Measuring Accuracy in a Post-Enumeration Survey." *Survey Methodology*, 14:1 99-116.

Isaki, Cary, Schultz, Linda, and Diffendal, Gregg, (1987), "Small Area Estimation Research for Census Undercount - Progress Report," pp 219-238, in *Small Area Statistics - An International Symposium,* R. Platek J.N.K. Rao, C.E. Sarndal and N.P. Singh, (eds) John Wiley and Sons.

Jaro, Matthew, 1989, "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida, *Journal of the American Statistical Association,* June, vol 84, No. 406, pp 414-420.

Sekar, Chandra and Deming, Edwards, W, (1949), "On a Method of Estimating Birth and Death Rates and the Extent of Registration." Journal of the American Statistical Association, 44, 101-115.

Winkler, William, and Thibaudeau, Y. 1990, "An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census," SRD Technical Report Series, Bureau of the Census, Washington, D.C. 20233.

Wolter, Kirk M, (1986), "Some Coverage Error Models for Census Data," *Journal of the American Statistical Association*, 81, 338-346.

Woltman, Henry, Alberti, Nicholas, and Moriarity, Chris, 1989, "Sample Design for the 1990 Census Post Enumeration Survey, "*Proceedings of the Survey Methodology Section of the American Statistical Association Annual Meetings.*

*This paper reports the general results of research undertaken by the Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.