Survey Methodology, June 1988
Vol. 14, No. 1, pp. 99–116
Statistics Canada

Measuring Accuracy in a Post-Enumeration Survey

HOWARD HOGAN and KIRK WOLTER[1]

ABSTRACT

The U.S. Bureau of the Census will use a post-enumeration survey to measure the coverage of the 1990 Decennial Census. The Census Bureau has developed and tested new procedures aimed at increasing the accuracy of the survey. This paper describes the new methods. It discusses the categories of error that occur in a post-enumeration survey and means of evaluation to determine that the results are accurate. The new methods and the evaluation of the methods are discussed in the context of a recent test post-enumeration survey.

KEY WORDS: Census; Undercount; Overcount; Coverage Evaluation.

1. INTRODUCTION

In this article we discuss recent research at the U.S. Bureau of the Census to improve the accuracy of a post-enumeration survey and to measure that accuracy. Much of this research was originally directed toward the goal of developing a sound body of statistical theory, methods, and operations for correcting U.S. census figures for coverage errors. The results presented in this paper show that we are now able to produce PES estimates of total population that are closer to the true population than are original census estimates.

In light of a policy decision made by the U.S. Department of Commerce not to correct the 1990 enumeration for coverage error, the PES methods we discuss will be used to provide a careful evaluation of the coverage of the 1990 Census. See U.S. Department of Commerce (1987). This evaluation will be used to inform users of the limitations of the census, to inform planning for future censuses, or to improve the Census Bureau's estimates of the U.S. population for years subsequent to the census year.

The PES method uses two samples to measure net coverage error. A sample of people who should have been counted in the original census enumeration is interviewed after the census and is used to measure census omissions. We call this the population or "P" sample. One also needs a sample of census enumerations to measure duplicates and other errors included in the census count. We call this the enumeration or "E" sample. The samples form an estimate of total population using the dual system-estimator (DSE). See Diffendal (1988) for a full discussion of the samples and the dual-system model. Unless otherwise stated, we will use Diffendal's notation throughout this article.

The Census Bureau conducted a PES in conjunction with the 1980 Census. The P sample consisted of persons in households enumerated in the April and August Current Population Survey (CPS) samples. For a description of the CPS, see U.S. Bureau of the Census (1978). The E sample was a separate and independent sample of persons in housing units enumerated in the census. In addition, the Census Bureau produced an alternative set of undercount estimates based upon an aggregate analysis of birth and death registration data, administrative

[1] U.S. Bureau of the Census, Washington D.C. 20233. Howard Hogan is Chief of the Undercount Research Staff. Kirk Wolter is Chief of the Statistical Research Division. This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributed to the authors and do not necessarily reflect those of the Census Bureau.

records, and previous censuses. This program, called demographic analysis, will be referred to occasionally in this article. The Census Bureau did not correct the 1980 enumeration for undercount errors because we considered the PES estimates to be flawed by missing and inaccurate data. In addition, the demographic analysis results were flawed by, among other things, a lack of data on the number of undocumented immigrants and the lack of an acceptable method to carry the estimates down to the state and local level. See Fay *et al.* (1988).

In very recent years, we have developed a new PES design and new methodology that minimizes the problems experienced in 1980, while not creating major new ones. The new PES design is based on a common area sample of census blocks for both the P and E samples. The P sample consists of all people living in the sample blocks at the time of PES interviewing. Interviewers visit each housing unit and determine where the residents were living at the time of the census.

Using newly developed computer matching methods and software (Jaro 1988), we attempt to match all P-sample people to corresponding census enumerations. Clerks review the computer's work and make a final determination as to the enumeration status (either enumerated or missed in the original enumeration) of each P-sample person. For people who moved between the census and the PES, we assign the census-day address to the proper block and search for a match there. For a few cases, matching is indeterminate at this point, and a further interview or followup is necessary either to gather additional information or to resolve conflicts in existing information. After the followup, clerks assign an enumeration status to the P-sample people for whom the followup interview is complete. For a very few residual cases, matching may be still unresolved, and we impute to each an enumeration status, using appropriate statistical techniques for missing data (Schenker 1988).

For each E-sample person, a determination is made as to the person's enumeration status (either correctly enumerated or erroneously enumerated) in the original census. Section 6 gives a description of what constitutes an erroneous enumeration (EE), and all non-erroneous enumerations are considered correct enumerations (CE). In many cases, the census enumerates the same people that are interviewed as part of the P sample. Thus, the two samples overlap to a great extent. Most E-sample people who are also in the P sample (as determined by the computer and clerical matching system) are automatically declared CE. However, the overlap is not complete. The P sample will miss some people that are included in the E sample and vice versa. The census will enumerate others in the block by mistake. Interviewers will invent some enumerations. For all E-sample people who are not matched to a P-sample person, it is necessary to conduct a followup interview. This followup gathers enough information to allow a determination of whether the E-sample people were counted correctly in the original enumeration.

We tested the new PES design in 1986 in connection with a test census in Los Angeles. The test was called the Test of Adjustment Related Operations (TARO) and consisted of 190 blocks, containing almost six thousand housing units and 20,000 people. The estimated net undercount for the Los Angeles test was about 9 percent. For details on TARO methods and results, see Diffendal (1988) and Schenker (1988).

We also tested the new PES design in a rural area of Mississippi during 1986. There we used a sample of 271 blocks with about 3250 housing units and eight thousand people. The estimated undercount in this test was 5.5 percent. For details of results and methodology, see Anolik (1988). Although, the Mississippi test data have not been as completely analyzed as the TARO data, we will refer occasionally to the results in this article.

An important question is whether the new PES can produce more accurate estimates of population than can the original census enumeration. In theory, the PES estimates should be considered the more accurate, but in practice, nonsampling errors can and do arise in the

Table 1

TARO Errors and Estimates of the Mean Effect on the Estimated Undercount of Correcting the Error

Sources of Error	Mean Effect on Estimated Undercount
Matching error	– 1.0%
Reporting census-day address	– 1.0%
Fabrication in the PES interview	– 1.0%
Missing data	0.0%
Error in measuring the erroneous enumerations	– 0.5%
Balancing gross overcounts and undercounts	0.0%
Correlation bias	+ 2.3%
Random error	0.0%

conduct and analysis of both the PES and the census enumeration. Careful study is needed to assess their relative accuracies. In this article, we present our assessment of the error structure of the 1986 TARO.

Eight potential sources of error affect coverage measurements produced by the PES: sampling error plus seven sources of nonsampling error. The sources and our summary assessment of their impact on TARO data are presented in Table 1. The second column gives the effects of the errors on the estimated undercount. For example, if we correct all "matching errors," the estimated undercount would be reduced by about one percentage point, from 9 percent to 8 percent. Some errors, such as "missing data" and "random error", might either raise or lower the undercount, and our best assessment is that these errors introduce no important bias into TARO data. The figures in this column represent assessments of individual error, without regard for the other sources of error.

By construction, the eight individual errors tend to be mutually exclusive and additive. Some overlaps or interactions are possible between the different sources, but we believe they are small and we ignore them here. Overall, we calculate the joint effect of the errors as

$$(-1.0 - 1.0 - 1.0 + 0.0 - 0.5 + 0.0 + 2.3 + 0.0) \text{ percent} = -1.2 \text{ percent}.$$

Thus, correcting for the joint effect of the errors would lower the estimated undercount from 9.0 percent to about 7.8 percent. The corrected figure, 7.8 percent, may be viewed approximately as the mean of a posterior error distribution for the TARO undercount. Development of a complete posterior error distribution is proceeding at the Census Bureau (see Mulry and Spencer 1988).

Because the original TARO estimate of 9 percent is much closer to the corrected figure of 7.8 percent than the corrected figure is to zero, we conclude that the original TARO data is closer to the truth than is the original census enumeration.

In the next 8 sections of the article, we treat the error components one by one. Each section discusses both the procedures and problems confronted in the 1980 PES, and the error-resistant improvements that were tested in TARO. We describe the evaluation of each error component and the evidence for our conclusions. The paper closes in Section 10 with a summary of our findings and some directions for future research.

2. MATCHING ERROR

Errors in classifying P-sample people as enumerated or not can occur for two general reasons:

(a) the information reported by the respondent/interviewer is incorrect
(b) correct information is reported, but not correctly used.

Category (a) consists of errors in the reporting of census-day address and fabrication in the PES interview, discussed in Sections 3 and 4, respectively. The present section discusses matching errors (category (b)) that occur even when the people are real and their çensus-day address is correctly reported. In other words, these are errors in matching due to processing mistakes.

In our new PES design, matching takes two forms: automated batch matching and computer-assisted clerical matching. The status of "not enumerated" is assigned to a P-sample person when sufficient information for matching has been gathered and no matching case can be found in the census. Errors occur when there actually was insufficient information for matching but matching was attempted nonetheless, and also when the correct census questionnaires were searched but the match was not established, even though the person was in fact counted in the original enumeration.

A P-sample person occasionally may be declared to match the wrong census person. This happens most often within families, where children's names and ages may be similar, and in "ethnic" neighborhoods where certain names are unusually common. Normally, false matches are less common than false nonmatches because the matches can be reviewed easily by a clerical matching staff. False matches create a bias in the dual system estimator only when the P-sample person was actually not enumerated.

A principal change in our PES design since 1980 that allows better control of matching error is the use of a common sample of blocks for both P and E samples. The block sample design permits a classification of all enumerated people (both P-and E-sample) into three categories:

— counted in P sample, counted in E sample
— counted in P sample, missing from E sample
— missing from P sample, counted in E sample.

This kind of organization or accounting, which was not possible with the 1980 design, imparts to the matching process a quality that resists matching error. For example, people with similar names in ethnic neighborhoods can be sorted out using all the information provided by a block sample. Address mix-ups in the census process are easier to handle with a block sample. The choice of census block as a sampling unit also reduces geographic coding error as compared to the 1980 PES, where the P sample was based on CPS clusters of four housing units and 1970 Census geography.

Matching is especially difficult for P-sample people who lived elsewhere on census day, i.e., movers. For movers, the census-day address reported in the P-sample interview must be assigned to the proper geographical area prior to matching. This assignment was problematic in the 1980 PES and the new design does not necessarily solve the problem. The Census Bureau will, however, be using a new, automated geographical system for the 1990 Census (see Marx and Saalfeld 1988), and we are hopeful that this innovation will permit rapid and accurate geographic assignment for mover addresses.

In the 1986 TARO, about 74 percent of the P-sample people were matched by the computer. Another 12 percent were declared "possible match" by the computer. A specially trained clerical staff reviewed all cases not designated as "match" by the computer, including all of the computer-designated "possible matches."

Table 2

Results of Rematch Study: Sample (Weighted)[a]

Results of Original Matching	Results of Rematching			
	Enumerated	Not Enumerated	Unresolved	Total
Enumerated	16,623	18	55	16,696
Not Enumerated	88	2,164	56	2,308
Unresolved	17	0	132	149
Total	16,728	2,182	243	19,153

[a] Weighting is to P-sample totals.

The results of the 1986 PES in Mississippi show that the success of the computer matching system is not limited to urban areas with house numbers, street names and well-defined geography. In the Mississippi test, addresses commonly consisted of a rural route and box number. Blocks were irregularly shaped with invisible boundaries such as an intermittent stream or county line. Still, the computer was able to match 68 percent of the cases.

We have conducted two studies to evaluate the extent of matching error in TARO. In the first study, a subsample of 35 blocks was selected and rematched by professionals from headquarters. The rematch was done independently of the original match, and then discrepancies between the match and rematch results were adjudicated. Because of this intensive approach to the rematch, we believe the rematch results represent true match status, while differences between the match and rematch results represent the bias in the original match results. Only nonmovers were considered in this study. Also, the study was confined only to within-block rematching, and thus did not formally measure any false nonmatches that may have occurred because the census enumeration was located outside the PES block.

The results for the P sample are given in Table 2 in the form of a cross-tabulation of match statuses as assigned from the original TARO match and the rematch.

We estimate there are about 88 false nonmatches and 18 false matches in the original TARO results, and that 111 = 55 + 56 cases originally matched or not matched should have been declared to have an indeterminate or unresolved match status. In the normal course of estimation, the unresolved would be treated by missing data procedures (Schenker 1988). The net result is that the observed match rate, i.e., the number matched divided by the number matched plus not matched, is .879 in the original match and .885 in the rematch, and thus that the original match rate is biased downward by about 0.6 percent.

The second evaluation study looked at the extent of matching error for movers. Among the original "not matched," there were 90 persons who reported moving between census-day and the time of the PES. For movers, searching is done at the reported census-day address. As an evaluation of the accuracy of the matching process, we reworked all 90 nonmatched mover cases using more intensive procedures. Eleven new matches were discovered, and as a result, the observed match rate for in-scope movers increased by .058, from .661 to .719. Although, the false nonmatch rate, 11/90 = .122, for movers is larger than we observed for nonmovers, the movers comprise a relatively small portion of the overall P sample. Correcting the 0.6 percent and 5.8 percent downward bias in match rate for nonmovers and movers has the overall effect of reducing the TARO undercount rate by 0.7 percent.

These calculations ignore the possibility of further new matches that might have been observed had the rematch study extended beyond the bounds of the PES blocks (Thompson, Whitford and Stoudt 1987). Based on evidence from computer matching across the Los Angeles test site, however, we conclude that geographical assignment was accurate, and that the incremental effect of such additional matches could do no more than to reduce the estimated TARO undercount by a further 0.3 percent.

3. REPORTING CENSUS-DAY ADDRESS

In our new PES design, as in the 1980 design, we attempt to match the P-sample people to the census enumeration at the census-day address. To facilitate the matching, the P-sample interviewer must ask where each household member lived on census day. The interviewer then probes for other addresses where the persons may have lived, including such places as at college or university, on a military base or ship, or at a second home. If the census-day address is reported incorrectly in the P-sample interview, then we may falsely designate the household members as not enumerated in the census, thus biasing upwards the estimated undercount rate.

To study address misreporting, we reinterviewed a subsample of the matched and unmatched cases after the original TARO estimates of undercount had been produced. This followup was six months after the initial PES interview and ten months after census day. Before presenting the results, we mention two limitations on this study. The first is the potential of greater recall error than in the original P-sample interview. Second, any trust created by the census advertising program may have faded, a potentially serious problem in an area with a large number of undocumented immigrants who fear all contacts with the government.

Table 3 describes the composition of the subsample. In most cases, the PES household matches the census household completely (''whole-household matches''). In the category ''partial-household matches,'' some of the PES persons match the census, but others do not. The ''whole-household nonmatch with conflicts'' category constitutes what we call the ''Emerson-Peterson'' problem. The census enumerated the ''Emersons'' at a particular address and the E-sample followup confirmed the census enumeration as correct. However, the P-sample interview showed the ''Petersons'' as living at the address on census day. These facts are in conflict, and one possible explanation is that the Petersons misreported their census-day address. The ''whole household nonmatches without conflicts'' category has no apparent contradictions; for example, the census missed the housing unit or listed it as vacant.

Table 3

Post-Production Followup Sample Sizes

Status of Original Match	Number of Households	
	Total in P sample	Rein-terviewed
Whole-Household Match	4,662	50
Partial-Household Match	609	50
Whole-Household Nonmatch with Conflicts	160	64
Whole-Household Nonmatch without Conflicts	357	109

Table 4

Outcome of
Post Production Followup, (Persons) Unweighted

Outcome	Whole-Household NonMatch				Partial-Household Match				Whole-Household Match	
	with Conflict		without Conflict		Non-matched		Matched			
	#	%	#	%	#	%	#	%	#	%
Address Confirmed	64	33	252	73	61	75	138	90	164	99
New Address Given	32	17	46	13	13	16	15	10	1	1
Possible Fabrication	70	36	23	7	2	2	0	0	0	0
Noninterview	27	14	24	7	5	6	0	0	0	0
Total	193	100	345	100	81	100	153	100	165	100

Note: # signifies number of people in the followup subsample.
% signifies percent of column category.

Table 4 gives the results for persons in the sample, with a separate breakout of initially matched v. nonmatched persons in partially matched households. As expected, the rates at which the address was confirmed vary greatly across strata. Virtually all addresses were confirmed for the persons in the whole-household match category, while the lowest rate of confirmation was for the whole-household nonmatch with conflicts category. New addresses were given by 13 to 17 percent of the nonmatched people across each of the three categories. Interestingly, new addresses were reported for ten percent of the matched people within partially matched households, not much less than for the nonmatched people within these households. The newly reported address is unlikely to be correct, unless identical errors were made in the original P sample and census interviews. This variable reporting reinforces our view that followup interviewing months after the original P-sample interview sometimes gives a different response (because of recall error and fear), but not necessarily a more accurate address.

Evidence was gathered on 95 cases that suggest they were possibly fabricated in the original P-sample interview. Most of these cases (70) came from the category of whole household nonmatches with conflicts. This problem is discussed further in Section 4. In addition, there were cases where the reinterview was not complete or yielded insufficient information to classify individuals into one of the categories. Some of these, had they been correctly interviewed, may also have reported a new address.

Weighting Table 4 to P-sample totals, we estimate that 3.1 percent of P-sample persons were erroneously reported as nonmovers in the original P-sample interview. For those who moved within the test site, we were able to search for a match at the new address, and we found that one third of those cases were enumerated in the Los Angeles test census. To assess the probable effect of reporting errors, particularly as we view TARO as a test of a national PES, we assume that those people who reported addresses outside the site would have been enumerated at the same rate as those who reported addresses within. Thus, one-third of the 3.1 percent would have been matched and classified as enumerated. Correcting for the reporting error results in a one percent reduction in the estimated undercount.

4. FABRICATION IN THE PES INTERVIEW

In spite of all good efforts to train and control interviewers, a PES interviewer may occasionally fabricate a household in lieu of conducting a proper interview. Fabricated cases will not match to the census. The estimated undercount rate will be inflated to the extent that fabricated cases substitute for people at the address who were actually enumerated.

Our new PES design seeks to control the fabrication rate to low levels. The sample design allows for frequent quality control checks using re-interviews of the interviewers' work. Samples are checked for each interviewer's work from each block several times per week. This close review was not possible in the 1980 PES, where interview assignments were not as highly clustered. We have also improved the training and supervision of the interviewing since 1980. Feedback on performance and retraining is now available to interviewers so that errors will not be repeated.

Two studies shed light on the extent of fabrication in the 1986 TARO. First, extensive quality control checks were performed during data collection for the P sample, both for address listing and for interviewing. The main conclusion from the quality control results is that there was evidence of only a small amount of fabrication. A total of 2070 P-sample interviews were checked by quality control clerks a few days after the original interview to verify the household composition (roster check). Of these, 59 interviews failed the roster check. These cases were examined in detail to determine how many of them were examples of fabrication. This was determined by whether each person in the household, as reported by the original interviewer (not the quality control clerk), matched to the census, which implies that the original interviewer collected valid data for that person. A clone fabrication in the census would be needed to invalidate this assumption. Only 13 of the 59 cases were identified as possible fabrications in that they had, for example, no persons from the original PES roster matching the census. Hence, the estimated fabrication rate for the quality control check is 0.6 percent.

The second source of data on the extent of fabrication is the post-production followup described in Section 3. From the data in Table 4, we estimate that about 1.2 percent of the P-sample people may have been obtained in fabricated interviews. This fabrication rate is about twice as large as provided by the quality control roster check. We believe much of the difference is attributable to one bad interviewer whose work was discovered in the followup interview, but evidently escaped detection by the quality control system. Another part of the difference may be that the followup exaggerates the level of fabrication; that is, landlords and other respondents deny the existence of people who occupy illegally converted housing units or who are present in the country without documentation.

To calculate an upper bound on the effect of fabrication in TARO, we assume the higher fabrication rate, .012, and we assume that if proper interviews had been conducted, the resulting P-sample people would match to the census at the same rate as achieved for the nonfabricated cases, or about .88. This leads to a corrected undercount of about 7.9 percent, about 1.1 percent lower than the original undercount of 9 percent. If we assume the lower fabrication rate, .006, then by similar calculations, the corrected undercount is 8.4 percent, or about .6 percent lower than the original TARO figure. In the summary of TARO errors presented in Table 1, we specified a value of 1 percent, which is about equal to the effect implied by the upper bound.

5. MISSING DATA

In order to measure small coverage errors accurately, the PES data set should be as complete as possible, without a large percentage of missing data. Unfortunately, there was a very large amount of missing data in the 1980 study (Fay *et al.* 1988). A number of changes in the PES design should now lead to lower levels of missingness.

Table 5

PES Missing-Data Rates (%)

Source	1980 PEP		1986 TARO
	April	August	
P Sample			
Noninterview (Household)	4.4	5.3	0.5
Unresolved enumeration status (Person)	4.0	4.4	0.8
Total	8.4	9.7	1.3
Proxy interview (Household)	a	a	3.2
E Sample			
Noninterview (Household)	1.1	1.1	NA
Geocoding indeterminate (Household)	1.6	1.6	NA
Unresolved enumeration status (Person)	2.0	2.0	4.7
Total	4.7	4.7	4.7

[a] Percent unknown.
NOTE: NA signifies "not applicable."

First, because of the tight time schedule for CPS interviewing, the initial P-sample interviews in 1980 were conducted during a one-week period. For the new PES, a three-week interviewing period is used, with yet another week if special problems arise. The longer interviewing period decreases the household noninterview rate. Another change that reduces the household noninterview rate is the sample of blocks (rather than list-sample clusters of four housing units as in the CPS). This sample allows the interviewer to visit a housing unit several times (perhaps between visits to the other housing units in the block) without extreme travel costs.

Incomplete followup interviews caused a large portion of the missing P-sample enumeration statuses in the 1980 PES (2.6 percent for April and 2.8 percent for August). We are attempting to diminish this problem by collecting the information needed to declare cases as either enumerated or missed during the initial interview, thereby eliminating the need for followup in most cases. Additionally, improvements in the timing and quality of matching, because of the new automated matcher, will reduce the number of cases requiring followup.

In the new PES design, the P and E samples overlap, and thus most of the information needed to determine E-sample enumeration statuses is gathered early, during initial P-sample interviewing. The use of a block sample, along with improved census geography, also helps reduce the proportion of E-sample cases for which correctness of census geocoding cannot be determined. Finally, improvements have been made in the treatment of missing data (Schenker 1988).

As can be seen in Table 5, the missing-data rates for the P sample in TARO are much lower than those for the 1980 PES. The E-sample total missing-data rate for TARO is equal to that for the 1980 PES, but this was due to an operational error in TARO, and we expect reductions in missing data similar to those for the P sample in the future.

Even though TARO achieved low levels of missing data, it is important to examine what effect the missing data has on the estimated undercount rates. To answer this question, we produced several sets of undercount estimates for TARO derived using alternative treatments of missing data, P-sample proxy interviews, P-sample movers, and certain E-sample unresolved cases. See Schenker (1988) for a detailed description of the alternative estimates, which ranged from a low of 7.8 percent to a high of 9.4 percent. Two of the alternative treatments considered in Schenker (1988) deal with problems discussed elsewhere in our paper; they are the treatment of movers within the test site (Sections 2 and 3) and E-sample resolved cases that may have been fictitious enumerations (Section 6). The effects of these treatments are attributed in Table 1 to sources of error other than missing data, and are the main reason for the difference between the TARO undercount estimate of 9 percent and the lowest alternative estimate of 7.8 percent. When the other treatments discussed in Schenker (1988) are considered, the change in the estimated undercount ranges from −0.3 percent to 0.3 percent. These changes are quite small and it is uncertain in which direction the true effect lies. Hence, we have listed a mean effect of 0.0 percent in Table 1.

6. ERROR IN MEASURING THE ERRONEOUS ENUMERATIONS

To estimate net coverage error, it is necessary to estimate the number of erroneous enumerations (EE) contained within the original census enumeration. EE includes the following distinct categories:

(i) fabrication in the census, where the census enumerator or respondent creates fictitious people in lieu of conducting a proper interview;
(ii) census duplicates;
(iii) persons born after census-day and persons who died before census-day; and
(iv) persons enumerated in the census with such sparse or incomplete information as to render them unmatchable to the PES.

All of these categories are estimated by way of the E sample. In addition, certain census geographic coding errors are treated as erroneous enumerations; this problem is part of the balancing issue discussed in Section 7.

In the 1980 PES, the E sample was a separate and independent sample of 110,000 census household enumerations. Interviewers revisited the housing units 8 months after census day to verify that the census enumerations were either correct or erroneous. Also, the housing unit was located on a map to see if it was assigned to the correct census geography, and clerks searched the census records to identify duplicates.

We have instituted two important changes in the new E-sample design. First, as already discussed, both the E and P samples will now be based on the same sample of blocks. We have found that overlapping P and E samples reduces geographic assignment errors. Second, most E-sample data will be collected in July, just three months after census-day. The procedures are such that most E-sample people are automatically designated correctly enumerated if they are counted in the P sample in July and are subsequently matched correctly to the person's E-sample enumeration. Unmatched E-sample cases are tagged for a followup interview, occurring only 6 months after census day. The earlier reporting in this new design lowers the missing data rates, reduces reliance upon proxy respondents, and improves the quality of the collected data.

There are four main components of error in the measurement of EE:

(i) response errors in the E-sample interview (this is the P-sample interview for most cases and the followup interview for all other cases), or mis-coding of responses by the processing staff;
(ii) error committed by an interviewer or by staff in assigning the correct geographic code to an E-sample person;
(iii) error in conducting the search for duplicates; and
(iv) mistakes made in classifying an E-sample case as having insufficient information for matching.

In addition, there are errors due to non-response in the E-sample interview, as discussed in Section 5, and sampling error, as discussed in Section 9.

Response errors often relate to the assignment of the status of "fictitious" to an E-sample person. The E-sample interviewer sometimes finds that the current resident of a unit (or another eligible respondent) does not know the people listed in the census. Usually, this is because the current resident moved in after the census and simply does not know who was living there at the time of the census. These E-sample cases should be designated as nonresponse. However, if the census enumerations were fabricated, no respondent will know the "people" reported in the census.

In experimenting with the new design in the TARO, the E-sample interviewers were instructed to determine whether the E-sample enumerations were fictitious and to record the basis for their decisions. Initially, the clerks required very strong evidence before designating an E-sample person as fictitious. It was this data that was used in preparing the first TARO estimates of total population and percent undercount. We realized that the rules for coding were being interpreted too strictly, and later, we had professionals review all E-sample cases coded as "noninterview, respondent does not know" to determine if any should have been coded as "fictitious". Out of 257 such E-sample cases, 118 were coded by the professionals as "fictitious." The corrected information was used to create some alternative TARO estimates (Schenker 1988).

Geographic assignment of census returns was generally thought to be very good in the Los Angeles test site, which was a long-established neighborhood with large well-defined blocks. We have not produced formal measures of the effects of geographic misassignment on the estimated EE, but we believe such error is negligible. In other areas of the U.S., however, the errors could be nonnegligible either because of poor maps, poor or incomplete addresses, or confusion about geographic locations created by new construction.

For example, in contrast with Los Angeles, geographic assignment was a problem in the 1986 Mississippi census returns. There we discovered 2.22 percent of the E sample was duplicated. Of the duplicate cases, 35 percent were located outside the sample block. Although we were able to find many duplicates outside the sample block, we are not convinced we found all of them. This is because searching for duplicates was not designed as a separate activity. We only identified duplicates in the course of other PES operations, and thus probably missed many of them. In the next PES, we will implement a separate activity to search for duplicates.

The census sometimes enumerates people with such sparse information that even if they were correctly interviewed in the P sample, a match to the E sample would not be possible. To compensate for this problem, such E-sample cases should be included in EE so as to estimate the total population properly. This problem is similar to that of geographic balancing discussed in Section 7. The separate E and P samples in the 1980 PES made it very difficult to do this consistently; similar cases were classified as "unmatchable" in the E sample and "matchable" in the P sample, thus creating a bias in the dual system estimator. Because the new PES design uses overlapping P and E samples, we ensure that identical rules are applied, thus eliminating the bias.

In another evaluation of the TARO, and as part of the rematch study discussed earlier (see Section 2), the E-sample cases in a subsample of 35 blocks were reprocessed by professionals from headquarters. As in Section 2, the rematch was independent of the original work, with subsequent adjudication of any discrepancies. Thus, we believe the rematch represents the best possible determination of the true enumeration statuses of the E-sample people, while differences between the original work and the rematch may be regarded as a measure of bias due to error in the original work.

Table 6

Results of Rematch Study: E Sample (Weighted)[a]

Original Results	Results of Rematching			
	Correct Enumeration	Erroneous Enumeration	Unresolved	Total
Correct Enumeration	19,153	28	88	19,269
Erroneous Enumeration	41	283	1	325
Unresolved	140	100	223	463
Total	19,334	411	312	20,057

[a] Weighting is to E-sample totals.

Results are presented in Table 6. Notice that most of the changes involve cases originally classified as "unresolved." Many of these cases were those discussed earlier, requiring a subjective decision between "fictitious" and "nonresponse." Based on these data, we believe that better clerical procedures are needed for coding E-sample cases as fictitious. We are presently working to implement improved procedures in the Census Bureau's next PES, to be done in conjunction with a 1988 dress rehearsal of the 1990 Census.

From the rematch study, we believe the original rate of EE,

$$\frac{325}{325 + 19{,}269} = .016$$

should be increased to about

$$\frac{411}{411 + 19{,}334} = .021 .$$

This implies the original TARO undercount should be reduced by about 0.5 percent. The corrected undercount is thus about 8.5 percent.

7. BALANCING GROSS OVERCOUNTS AND UNDERCOUNTS

In order to estimate net undercoverage, the methods and concepts used to measure gross overcount must be consistent with those used to measure gross undercount. We refer to this requirement as "balancing." We proceed to give an elementary description of how the PES achieves balance.

One way to view this issue is to consider the dual system estimator in the form

$$\hat{N}_{++} = (\hat{N}_{1+}\hat{N}_{+1})/\hat{N}_{11},$$

where

$$\hat{N}_{11} = M,$$

the weighted number of matched P-sample people, and

$$\hat{N}_{+1} = N_p,$$

the weighted number of people in the P sample. All notation is defined in Diffendal (1988).

Since we cannot search all census questionnaires, the observed number matched, M, will be lower than the true number in both systems. To make costs manageable, matching for a given case is restricted to a "search area", typically the sample block and one or two rings of surrounding blocks.

As a consequence, the term $\hat{N}_{11}$ estimates kN^*_{11}, where $0 \leq k \leq 1$ is the conditional probability that a census enumerated individual is counted in the correct search area and N^*_{11} is the PES estimator of N_{11} that would obtain if it were feasible to conduct the search for matches over the entire population.

To construct a consistent estimator of population size, we must reduce the number counted in the census by the factor k. Because the E-sample search for erroneous enumerations, e.g., duplicates, extends over the search area and we treat as erroneous all enumerations that should not be included in the search area, the term $\hat{N}_{1+}$ estimates $k\, N^*_{1+}$, where N^*_{1+} is the estimator of N_{1+} that would obtain if it were feasible to conduct the search for erroneous enumerations over the entire population.

Assuming consistent search areas, the DSE becomes a consistent estimator of N_{++}. Note that in this model of the balancing process, we do not estimate the probability k, but instead rely on consistent search areas to eliminate it from the DSE.

Balancing the P sample and the E sample in the 1980 PES was impossible because the samples did not overlap. The CPS (or P-sample) addresses were coded to census geography. The search area was to have been limited to a close neighborhood of the CPS address, but because the CPS addresses were based on 1970 Census geography, they could not be easily assigned 1980 Census geographic codes, and searching extended over a wide area. As the search area expanded for the P sample, the E-sample search area should also have expanded. We believe inconsistencies arose between E-and P-sample search areas, thus creating a bias in the DSE.

In TARO, we performed the two-way match between the P-and E-sample persons within the selected blocks. The geography and search areas were consistent, well-defined, and well-controlled during computer and clerical matching. As a consequence, the problem of balancing did not introduce any important bias into the Los Angeles results.

8. CORRELATION BIAS

For the dual system estimator to be a consistent estimator of the true population size N_{++}, two independence assumptions are needed:

(i) causal independence,

(ii) heterogeneous independence.

In addition, autonomous independence is often assumed, but failure of this assumption is known to impart little or no bias to the estimate of total population. (Wolter 1986b and Cowan and Malec 1986).

Causal independence fails when an individual's capture history in the census alters the probabilities of capture in the PES. The estimator $\hat{N}_{++}$ is downward biased when the odds of capture in the PES are increased as a result of capture in the census, and is upward-biased when the odds of capture in the PES are reduced as a result of capture in the census.

An important bias may exist in the April 1980 PES data because of a failure of causal independence. The failure occurred because respondents may have mistaken the April or March CPS enumerations for the census enumeration.

Table 7

Undercounts (%) for Black and Total Population the 1980, 1960 and 1950 U.S. Censuses, and Differential Undercount Rates

Source	Black	Total	Difference
1950			
PES	3.2	1.4	1.8
DA	9.6	4.4	5.2
1960			
PES	3.8	1.9	1.9
DA	8.3	3.3	5.0
1980			
PES[a]			
Low	1.1	- 1.0	2.1
Middle	6.9	1.4	5.5
High	5.7	2.1	3.6
DA	5.9	1.4	4.5

[a] The 1980 PES produced 12 sets of estimates. The three presented here are selected from the highest, middle and lowest set as measured by estimated total undercount.

Heterogeneous independence fails when census capture probabilities are different from one individual to another. The resulting bias (called heterogeneity bias or correlation bias) is generally thought to be a downward bias because individuals with a high probability of capture in the census also tend to have a high probability of capture in the PES and, conversely, individuals with a low probability of capture in the census also tend to have a low probability of capture in the PES.

Sekar and Deming (1949) suggested post-stratification to control heterogeneity bias. In practical applications, it is unlikely that this technique is fully effective; there is inevitably some residual heterogeneity of capture probabilities within post-strata.

In the dual-system model, the number of people missed by both systems, N_{22}, is estimated by

$$\hat{N}_{22} = \hat{N}_{12}\hat{N}_{21}/\hat{N}_{11},$$

as in Diffendal (1988), equation(2). Because the dual system estimator may be expressed in the form

$$\hat{N}_{++} = \hat{N}_{11} + \hat{N}_{12} + \hat{N}_{21} + \hat{N}_{22},$$

and because $\hat{N}_{11}$, $\hat{N}_{12}$, and $\hat{N}_{21}$ are direct design-based estimators, any bias due to failure of the independence assumptions arises solely in $\hat{N}_{22}$ as an estimator of N_{22}.

We can study the correlation bias in 1980 and previous censuses by comparing $\hat{N}_{++}$ to independent demographic analysis (DA) estimates of total population. Table 7 presents relevant data from recent censuses. If one treats demographic analysis estimates as a standard, these comparisons display total bias in the dual system estimator, including both correlation bias and other sources of error. We believe that the downward bias shown in these estimates is largely attributable to correlation bias. The 1950 PES gave severe underestimates of the population size, of the percent undercount, and of the differential undercount, presumably because of both causal and heterogeneity bias. Note, however, that if 1950 PES data had been used to correct the 1950 census, the differential undercount would have been reduced from 5.2 percentage points to approximately 3.4 percentage points.

The 1960 PES gave similar underestimates of population size, of the percent undercount and of the differential undercount, again presumably because of correlation bias. If the 1960 PES data had been used to correct the 1960 census, the differential undercount would have been reduced from 5.0 to approximately 3.1 percent.

No PES was conducted in 1970. The 1980 PES produced 12 sets of estimated undercounts based on the April and August results and on different sets of assumptions. The DA undercount rates are approximately in the middle of the 12 PES undercount rates. Correlation bias is not as evident here as in 1950 or 1960, largely because of improvements that were made in 1980 to reduce positive causal dependence. We believe the heterogeneity bias is still present but is obscured by other PES errors and by bias due to negative causal dependence.

In the new PES design, we attempt to control the bias due to causal effects by scheduling the PES enumeration after most major census field activities. This approach, contrary to that of the April 1980 PES, will promote causal independence between the census and PES enumerations as much as possible. Further, we are now using field office procedures that will promote causal independence, such as assigning PES interviewers to different areas than they worked (if they worked) in the original census enumeration.

It will be difficult to eliminate the correlation bias due to heterogeneity in future PES's. The only possible avenues include more effective post-stratification and combining the PES and DA data in some way, possibly by controlling for DA sex ratios. See Wolter (1986c) and Choi, Steel and Skinner (1988). We have done some experimentation this decade with alternative post-stratification schemes including using variables such as owner/renter status, census mail-back rate, and marital status. These approaches show some promise. See Diffendal (1988).

TARO yielded observed differential undercounts consistent with expected differentials. In the U.S., census coverage is normally lower for males than females. This result has been consistently observed from the results of demographic analysis. The TARO sex ratios (males per 100 females) are higher than the census ratios for Hispanics and for people who were neither Hispanic nor Asian. The TARO sex-ratios are much higher than census sex-ratios (1.1 to 3.4 more males per females) for the 30-44 year age group. This outcome is consistent with the 1980 national results from demographic analysis. Thus, we believe that the TARO sex ratios are closer to the true sex-ratios, and although correlation bias limits the gain, the PES is still able to measure the differential undercount.

Table 8 presents the two-way table of data for the 1986 TARO, with no post-stratification. The estimate of the number missed by both systems,

$$\hat{N}_{22} = 5{,}870,$$

is approximately the same order of magnitude as census substitutions 5,259 and erroneous enumerations 6,426. Approximately one-eighth of the estimated census misses, $\hat{N}_{12} + \hat{N}_{22} = 44{,}373$, are attributable to the (2,2) cell. Thus, most of the measured undercount arises from direct survey estimation, not from the dual-system model.

To illustrate the effect of correlation bias, consider doubling the size of the (2,2) cell. This increases the estimated undercount rate by about 1.4 percent. Based upon analysis of the 1980 PES, Ericksen and Kadane (1985) suggest multiplying the (2,2) cell by 2.7, thus increasing the estimated undercount by 2.3 percent.

We have other information that sheds light upon the problem of correlation bias. Three anthropologists worked for the Census Bureau as participant or systematic observers in the Los Angeles test. Their observations do not provide direct measurements of correlation bias, but rather they provide insights into the degree to which the census and PES are missing the

Table 8

Dual-System Estimates for 1986 Los Angeles Test Census

		PES		
		Counted	Missed	Total
Correct Census Enumerations[a]	Counted	298,204	45,463	343,667
	Missed	38,503	5,870	44,373
	Total	336,707	51,333	388,040

[a] Correct Census Enumerations = Total Census Enumerations − Substitutions − Erroneous Enumerations.

same kinds of people. The reports suggest that there are people with very low capture probabilities who tend to be missed by both the census and the PES, and thus that an important downward bias may be present in TARO data. See Hainer *et al.* (1988) and Hines (1988).

Given the data available, we have no exact means of assessing the level of correlation bias in the TARO data. Nevertheless, based upon the work just cited, we speculate that the TARO undercount rate may be too small by 2.3 percent or more.

9. RANDOM ERROR

Sampling error affects the estimates of the number of matches, the number of erroneous enumerations, and the P-sample totals. The census count and the number of substituted census people are based upon the 100 percent census enumeration, and as such are not contaminated by sampling error. The estimated standard deviation for the undercount rate is 0.007. So a 95 percent normal-theory confidence interval for the undercount rate is $.09 \pm 2\,(.007) = (.076, .104)$.

Diffendal (1988) presents estimated standard errors for the TARO adjustment factors defined by $Y = \hat{N}_{++}/CEN$ and used a components-of-variance model to smooth the Y, thus reducing the effects of sampling error. In most cases, the smoothing substantially reduced the estimated standard errors, particularly for domains. We believe such smoothing can be used profitably in future PES's.

10. CONCLUSION

After the 1980 Census, the Census Bureau reviewed its coverage measurement program and identified the program's weaknesses. We instituted a research program and a new coverage measurement design aimed at reducing the weaknesses. We have completed major tests of the new PES design this decade and have demonstrated substantial improvements over the 1980 PES.

In this article, we reviewed the results of our research program as reflected in the 1986 TARO. There may never be a perfect PES. However, none of the weaknesses or errors in the new design are so large as to invalidate the PES results. For reasons stated in Section 1, we believe the joint effect of the errors in the coverage measurement in TARO is smaller than the error in the original enumeration in Los Angeles.

One of the main benefits of the TARO is that it enables us to identify new questions and minor unresolved problems that warrant further research. For example, the initial PES interview attempted to gather the information needed to declare a P-sample person as missed in the census. We are now refining the questionnaire design, including additional screening questions to identify movers more accurately. In future PES's, we will also conduct followup interviews for most movers and for nonmover households in the P sample suspected of having misreported mover status. In this way, we believe mover misreporting can be kept to a minimum.

The quality control procedures that are intended to detect and correct fabrication in the PES must continue to be improved and tested. In addition to verifying names on the PES roster, other items shall be verified as part of the quality control check. This should detect any partial fabrication that occurs by obtaining names from mailboxes or landlords, and fabricating the characteristics. We are revising the PES followup forms in order to facilitate the identification of fictitious people.

Our goal for future PES's is to minimize missing data, especially through minimizing the need for followup. However, as more cases are sent to followup, the proportion of failed followup cases will increase. Research is needed on the proper treatment of these cases.

Not withstanding the good results from TARO, one should exercise appropriate caution before drawing the conclusion that the 1990 PES results will be closer to the truth than will the 1990 original enumeration. The actual level of net undercount in the Los Angeles test was high compared to what would be expected in a national census. Will the size of the errors in a national PES be small enough to produce more accurate population estimates?

We believe that the 1990 Census will contain areas with large undercounts and perhaps large overcounts, even if there is a small net national undercount. Thus, the PES should produce the more accurate population estimates for the areas most difficult to count. Through further polishing of the new PES during the last two years of this decade, it may be possible to produce more accurate population estimates for other, less-difficult-to-count areas too.

We also believe that the errors in the PES will decrease as the undercount decreases. Stable areas with good maps, well-defined addresses, few movers and cooperative respondents will be relatively easy for both the census and the PES. Residual processing errors may produce a threshold of accuracy beyond which the PES may not go, regardless of the true net undercount. We will not know for sure until the 1990 PES is executed. This situation may lead the PES estimates to be more accurate than original census estimates for some areas, with equal or nearly equal accuracy for most other areas. Statistical theory should provide a means to produce a best estimate by combining the results of the original enumeration and the PES.

ACKNOWLEDGEMENTS

The results presented represent the work of many people besides the authors, including Dan Childers, Carol Corby, Gregg Diffendal, Charisse Jeffries, Arona Pistiner, Nathaniel Schenker, Maria Urrutia, and Kirsten West. The authors would also like to thank the three referees for their many useful comments.

REFERENCES

ANOLIK, I. (1988). The rural post-enumeration survey in east central Mississippi. Statistical Research Division Report, Series RR 88/10. U.S. Bureau of the Census, Washington, D.C.

CHOI, C.Y., STEEL, D.G., and SKINNER, T.J. (1988). Adjusting the 1986 Australian Census for underenumeration. *Proceedings of the Census Bureau Fourth Annual Research Conference*. Bureau of the Census, Washington, D.C.

CITRO, C.F., and COHEN, M.L. (1985). *The Decennial Census: New Directions for Methodology in 1990*. Washington: National Academy Press.

COWAN, C.D. and MALEC, D. (1986). Capture-recapture models when both sources have clustered observations. *Journal of the American Statistical Association*, 81, 347-353.

DIFFENDAL, G. (1988). The 1986 test of adjustment related operations in central Los Angeles county. *Survey Methodology*, 14.

ERICKSEN, E.P., and KADANE, J. (1985). Estimating the population in a census year: 1980 and beyond. *Journal of the American Statistical Association*, 80, 98-114.

FAY, R.E., PASSEL, J.S., ROBINSON, G., and COWAN, C.D. (1988). The coverage of population in the 1980 Census. Technical report PHC 80-E4. Bureau of the Census, Washington, D.C.

HAINER, P., HINES, C., MARTIN, E., and SHAPIRO, G. (1988). Research on improving coverage in household surveys. *Proceedings of the Fourth Annual Research Conference*, Bureau of the Census, Washington, D.C.

HINES, C. (1988). The role of participant observation research in understanding the census undercount. Paper presented at the Population Association of America Annual Meetings, New Orleans, La.

JARO, M. (1988). Advances in record linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association* (forthcoming).

MARX, R.W. and SAALFELD, A.J. (1988). Programs for assuring map quality at the Bureau of the Census. *Proceedings of the Bureau of the Census Fourth Annual Research Conference*, Bureau of the Census, Washington, D.C.

MULRY, M., and SPENCER, B. (1988). Total error in dual system estimates of population size. *Proceedings of the Fourth Annual Research Conference*, Bureau of the Census, Washington, D.C.

SEKAR, C.C., and DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.

SCHENKER, N. (1988). Handling missing data in coverage estimation, with application to the 1986 test of adjustment related operations. *Survey Methodology*, 14.

THOMPSON, J.H., WHITFORD, D., and STOUDT, D. (1987). Memorandum for Howard Hogan, Subject: Review of 1986 PES Matching, April 21, 1987.

U.S. BUREAU OF THE CENSUS (1978). *The Current Population Survey: Design and Methodology*, Technical Paper No. 40, Washington, D.C.

U.S. DEPARTMENT OF COMMERCE (1987). Press notes, Statement by Undersecretary Robert Ortner, October 30, 1987.

WOLTER, K.M. (1986a). Some coverage error models for census data, *Journal of the American Statistical Association*, 81, 338-346.

WOLTER, K.M. (1986b). A combined coverage error model for individuals and housing units. Statistical Research Division Report Series RR 86/27, U.S. Bureau of the Census, Washington, D.C.

WOLTER, K.M. (1986c). Capture-recapture estimation in the presence of a known sex ratio. Statistical Research Division Report Series RR 86/20, U.S. Bureau of the Census, Washington, D.C.