Journal of Official Statistics
Vol. 7. No. 4, 1991, pp. 475–498

A Comparison of the Missing-Data Treatments in the Post-Enumeration Program

Joseph L. Schafer[1]

Abstract: Missing data were a major source of uncertainty in the U.S. Census Bureau's Post-Enumeration Program (PEP) to measure coverage error in the 1980 census. A variety of reweighting and imputation techniques led to widely different estimates of the undercount. This paper outlines the important features of missing data in the PEP undercount estimation, identifying why undercount estimates were sensitive to the various missing-data treatments. The twelve sets of PEP estimates should not be regarded as equally plausible alternatives, nor should their range be regarded as a measure of the uncertainty due to missing data. A distinct possibility exists that all of the missing-data treatments may have been inherently conservative, causing all of the estimated rates of gross undercount to be biased downward.

Key words: Dual-system estimation; ignorable nonresponse; imputation; undercount.

1. Introduction

In 1980, the U.S. Census Bureau conducted a large scale effort, known as the Post-Enumeration Program (PEP), to measure errors of coverage in the 1980 Decennial Census. The PEP used survey methods to measure both census undercounting and overcounting. Estimated population counts, corrected for net coverage error, were published for cities, states, and regions, and within categories of age, sex, and race (Fay, Passel, and Robinson 1988).

Missing data were a major source of uncertainty in PEP. Although the rates of missingness were not high by most survey standards, coverage estimates were sensitive to the procedures used to correct for missing data. Twelve sets of undercount estimates were published, reflecting a variety of alternative imputation and weighting adjustment procedures. The variability of estimates over the twelve sets led many to conclude that the uncertainty due to missing data alone was enough to make any of the estimates unreliable. The Census Bureau decided not to adjust any of its official 1980 census counts on the basis of PEP.

By this time, interest in the 1980 PEP is primarily academic; efforts to measure coverage error in the 1990 census are now

[1] Department of Statistics, Pennsylvania State University, University Park, PA 16802, U.S.A.

This research was supported in part by Joint Statistical Agreements 87–07, 88–02, and 89–08 between the U.S. Bureau of the Census and Harvard University, and in part by the Statistical Support Division, U.S. Bureau of the Census. This research was partially conducted while the author was employed by the Statistical Support Division, U.S. Bureau of the Census. The views expressed are attributable to the author and do not necessarily reflect those of the bureau. The author would like to thank Donald B. Rubin and Alan M. Zaslavsky of the Department of Statistics, Harvard University, for their many helpful comments and suggestions. Special thanks to Robert E. Fay of the U.S. Bureau of the Census for providing extensive written comments on an early draft of this paper.

complete, and attention is now focused on the 1990 results. Yet, a discussion of the PEP is useful because it touches upon three important issues regarding missing data that often arise in survey practice: (1) how viewing missing-data methods in terms of probability models helps to clarify the underlying assumptions and facilitate judgements about the merits of competing methods; (2) how missing-data procedures that do not make full use of the observed data can introduce substantial biases; and (3) the importance of conducting principled sensitivity analyses, which make best possible use of the data at hand, but vary those assumptions that are truly untestable from the data.

This paper outlines the important features of the missing data in the PEP, evaluates the various missing-data treatments, and discusses which of the PEP undercount estimates probably contain the smallest nonresponse bias. Some limited conclusions are drawn about where the best estimates of undercount obtainable from PEP might lie. The twelve sets of estimates should not be regarded as equally plausible alternatives, nor should the range of these estimates be regarded as a measure of the real level of uncertainty due to missing data. In fact, under very realistic assumptions about the missingness mechanisms in PEP, all of the estimates of gross undercount could well be biased downward.

For brevity, this paper deals exclusively with PEP's estimation of gross undercount, because the net coverage estimates were especially sensitive to the treatment of missing data for gross undercount. Missing data were also an important factor in the PEP estimation of gross overcount, however; many of the same issues arose in the missing-data treatments for gross overcount, and hence many similar comments would apply there as well.

Section 2 gives a brief overview of the coverage estimation methodology used in PEP. Section 3 discusses the sensitivity of coverage estimates to the various missing-data procedures. Sections 4 and 5 describe the types of missing data and their treatments. Criteria for evaluating the treatments are presented in Section 6, and the treatments are compared and analyzed in Section 7. Section 8 presents concluding remarks.

2. Overview of the PEP

2.1. Introduction

The PEP combined the technique of dual-system estimation, which measures gross undercount, with independent estimates of gross overcount, to measure net coverage error in the 1980 census (Fay, Passel, and Robinson 1988). This dual effort required the selection of two samples, the P-sample and the E-sample.

2.2. Estimating the undercount: the P-sample

To measure undercount, a sample of housing units, known as the P-sample, was selected and the persons within these housing units were enumerated shortly after Census Day, April 1, 1980. This enumeration was conducted in addition to and independently of the census. Census records were then searched clerically in an attempt to match the *P*-sample persons to census persons. The proportion of P-sample persons who could not be found in the census provided an estimate of the gross undercount rate.

The 1980 PEP actually used two P-samples, both derived from the Current Population Survey (CPS), a monthly survey conducted by the Census Bureau primarily to measure characteristics of the U.S. labor force. CPS interviewers used a supplemental questionnaire to collect additional information needed for the PEP at the end of the regular CPS interview. The first P-sample consisted of

the April 1980 CPS, which was conducted in mid-April, two to three weeks after Census Day. The second P-sample was taken in August, to provide an independent set of undercount estimates, and to assess the effect of an additional four-month time lag between the census and the P-sample on the undercount estimation procedures. Each was a nationwide multistage cluster sample of approximately 84,000 housing units. Details of the CPS sample design are given in Fay, Passel, and Robinson (1988).

Those persons who moved into P-sample housing units between Census Day and the time of the P-sample interviewing were considered to be part of the P-sample and were included in the undercount estimation. For those individuals, the census records had to be searched at their reported Census Day addresses to determine whether they were counted in the census. The special problems associated with these P-sample movers are discussed in Section 5.

2.3. *Estimating the overcount: the E-sample*

The census is subject to problems of overcounting as well as undercounting. Overcounts, or erroneous enumerations, include duplicates (persons enumerated in the census more than once), fictitious persons, and other types of definitional errors (persons not alive or living outside the country on Census Day, or who otherwise did not reside at the given housing unit according to census definitions of residency). All of these errors inflate the census counts. To estimate the number of such errors, a separate sample of 110,000 households, called the E-sample, was drawn from the 1980 census records. Reinterviews were attempted for the entire E-sample to verify whether the persons listed in the census for those units actually resided there on Census Day, according to census definitions of residency. For a 50% subsample of the E-sample, nearby census records were searched for duplicates, providing an estimate of the rate of duplication.

Another class of erroneous enumerations arises from geocoding errors, when persons are properly enumerated in the census but their housing units are incorrectly recorded in a different geographical area. Although geocoding errors do not affect population counts on the national level, they distort counts for small geographical areas, since every geocoding error creates an overcount in one place and an undercount in another. To estimate the number of geocoding errors, the physical location of each E-sample housing unit was verified to check whether the census geocoding was correct. Details of the E-sample operations are provided in Fay, Passel, and Robinson (1988).

2.4. *Dual-system estimation*

An estimate of the total population size may be obtained by dividing the census count by the P-sample estimate of the census coverage rate. This estimate, known as the dual-system estimate, may be written as

$$\hat{N} = \frac{N^C N^P}{M} \tag{1}$$

where $\hat{N}$ is the estimated population size, N^C is the census count, N^P is the total number of persons in the P-sample, and M is the number of P-sample persons also counted in the census (matched to the census records). This estimate assumes that the undercount rate among individuals in the P-sample and among individuals in the whole population are the same, except for the error of random sampling, i.e., that the P-sample is representative of the whole U.S. population with respect to undercount. Violation of this assumption is

known as correlation bias. A detailed discussion of the assumptions of dual-system estimation is given by Wolter (1986a).

A problem with the estimate (1) is that $\hat{N}$ can never be smaller than N^C; it corrects the census only for gross undercount. To correct for gross overcount as well, the estimated number of erroneous enumerations in the census should be subtracted from the census count N^C. Another deficiency in (1) is that it assumes census data are of sufficiently good quality that all individuals can be unambiguously identified from their recorded characteristics. In practice this is not the case; the census includes many persons for whom data were substantially incomplete, and whose characteristics were imputed. A P-sample person who was represented in the census, but for whom most of the identifying characteristics had been imputed, could never be conclusively matched to the census records. For consistency, then, the number of census persons who were not data-defined (i.e., who did not have enough identifying characteristics recorded in the census to allow a match to the P-sample) and whose census characteristics were imputed, should also be subtracted from N^C.

Correcting for erroneous enumerations and census imputations, the dual-system estimate is

$$\hat{N} = \frac{(N^C - I - E)N^P}{M} \qquad (2)$$

where I is the number of imputed persons in the census, and E is the number of erroneous census enumerations estimated from the E-sample. In PEP, dual-system estimates of the form (2) were calculated within poststratification cells defined by state, age, sex, and race, to increase precision and to reduce the effects of correlation bias. Weighted totals from the P and E-samples contributed the estimates N^P, M, and E, which were (approximately) design-unbiased.

3. Sensitivity of Coverage Estimates to Alternative Missing-Data Treatments

3.1. *Twelve sets of PEP estimates*

Due to its complex nature, the PEP was subject to a wide variety of nonsampling errors in addition to the ordinary variation of random sampling. Census coverage estimation was hampered by the poor quality of data provided by some respondents, inconsistent application of definitions and procedures, errors in the P and E-sample matching operations, violation of model assumptions, and missing data. Of the identified sources of nonsampling error in PEP, missing data were of primary importance.

Under a variety of alternative treatments for missing data, the Census Bureau produced five different P-sample datasets and three different E-sample datasets for dual-system estimation. By pairing the P and E-sample sets in all possible combinations, fifteen sets of dual-system estimates were possible, but results were published for only twelve. The twelve P/E combinations gave widely different estimates of the national undercount. They also gave widely different estimates of the differences in undercount between demographic subpopulations.

The estimated national undercount rates for the total U.S. population and for the black population under the twelve combinations are shown in Table 1. The highest national undercount rates, 1.9% overall and 7.8% for blacks, are implied by combination 2-20, which pairs P-sample Set 2 with E-sample Set 20. The lowest rates are implied by combination 14-8, which estimates an overcount of 1.0% nationally, and an undercount of only 1.1% for blacks. The variability of the estimates over the different

Table 1. Estimated 1980 national percent undercount rates for total noninstitutional population and black noninstitutional population by P-sample and E-sample set. Estimated standard errors due to sampling are 0.2 and 0.6 percent for the total and black populations, respectively, under all sets

P-Sample Set		E-Sample Set					
		Set 8		Set 9		Set 20	
		Total	Black	Total	Black	Total	Black
April	Set 2	1.1	6.1	1.5	7.3	1.9	7.8
	Set 3	1.0	5.7	1.4	6.9	1.7	7.4
	Set 14	−1.0	1.1	−0.5	2.3	−0.2	2.8
August	Set 5	1.7	4.5	2.1	5.7	—	—
	Set 10	0.3	2.8	—	—	—	—

Source: Fay, Passel, and Robinson (1988), table 7.1.

P-sample sets is much greater than the variability over the E-sample sets, partly because the rates of missing data in the P-samples were almost double that of the E-sample. In particular, a comparison of P-sample Set 14 with Set 2, within any E-sample set, shows a difference in estimated undercount rates of about 2% overall and 5% for blacks. All of these differences are much greater than the estimated standard errors due to sampling.

3.2. *Concern over the variability of the twelve sets*

The variability in these twelve sets of estimates was a major source of concern in interpreting the results of PEP. Firm conclusions about 1980 census coverage would be impossible if these twelve sets of estimates represented equally plausible alternatives. The sensitivity of coverage estimates to the missing-data treatments was undoubtedly a major factor in the Census Bureau's decision not to use PEP to adjust the official 1980 census counts; see, e.g., Bailar (1985) and Wolter (1986b). The American Statistical Association Technical Panel on the Census Undercount recommended that, in addition to providing alternative estimates to illustrate uncertainty, the Census Bureau should conduct the research necessary to produce a preferred set of estimates (ASA Technical Panel on the Census Undercount 1984), but no preferred set was chosen by the bureau.

The figures in Table 1 raise some important questions, including: Are these twelve sets of estimates equally plausible, or should one or more sets be preferred? Do they accurately reflect our range of uncertainty about census coverage? These issues have been repeatedly discussed in the statistical literature; see comments by Fay (1985) and Bailar (1985) accompanying the article of Ericksen and Kadane (1985), the discussions by Kadane (1986), Ericksen (1986), Madansky (1986), and the rejoinder by Freedman and Navidi (1986), and the article of Ericksen, Kadane, and Tukey (1989). The relative plausibility of the twelve sets was a major point of contention in *Cuomo v. Baldridge*, 80 Civ. 4550 (JES), in which the State of New York sought to enjoin the U.S. Department of Commerce to adjust the 1980 census counts for coverage error; see the affidavits of Ericksen (1983) and Wachter (1983), and the testimony of Cowan (paragraphs 630–631) and Stoto (paragraphs 712–714). A more detailed technical discussion is found in the exchange between Schafer (1988) and Fay (1989).

In the remainder of this paper, it will be

argued that P-sample Sets 2 and 5 are probably the most efficient and least biased; that Set 14 is difficult to justify either formally or intuitively, and is based on an implicit model that is quite extreme according to the best available evidence; and that the range of estimates under these twelve sets does not accurately reflect the uncertainty due to missing data.

4. Missing Data in the P-Sample

4.1. *Introduction*

The types of missing data that arose in PEP, and the procedures that were developed to handle these missing data, are detailed in Fay, Passel, and Robinson (1988), Fay (1988a, b). Some missing data were the result of clerical mishandling, lost materials, and unfortunate operational decisions which, in hindsight, could have been avoided. Other data were missing for reasons intrinsic to the PEP methodology, however, and would still have been missing under any realistic scenario. The latter types of missing data are the main focus of this discussion.

In the P-sample, the three major classes of missing data were household noninterviews, incomplete followups, and ungeocodeable cases. In addition, certain other classes of cases, though not missing data per se, were sometimes treated as missing or were omitted from the estimation as part of alternative P-sample treatments. Cases treated in this manner were April noninterviews whose data could be reconstructed from March and May, and August movers.

4.2. *P-sample noninterviews*

Of the approximately 84,000 households in the CPS sample, 4.4% were not successfully interviewed in April, and 5.3% were not successfully interviewed in August. For these units, household size and composition at the time of the P-sample interviewing were not known; the only information practically available concerned their location in the sample, i.e., stratum and cluster information.

4.3. *Incomplete followups*

In the initial round of clerical matching, 87% and 84% of the persons from interview households could be conclusively matched to the census in the April and August P-samples, respectively. (Unless otherwise noted, all P-sample figures quoted here are weighted by inverse probability of individual selection in the CPS, and are derived from tables in Fay (1988a)). The remaining persons could have been missed by the census, or they could have been unmatched for a variety of other reasons. One major reason, other than census omission, is that the correct Census Day address had not been searched in the matching operation.

Those that could not be matched were designated for followup interviews. The purpose of followup was to collect better information on Census Day addresses for the unmatched persons. Followup interviews were attempted early in 1981, roughly one year after the census. Many of these attempts were unsuccessful, however; 23% of the April and 19% of the August cases designated for followup did not have a complete followup interview. These incomplete followups remained unresolved with respect to their final match status (i.e., whether they were enumerated in the census or not).

4.4. *Ungeocodeable cases*

After followup interviewing, all cases whose followup interviews were successful were subjected to a final round of matching.

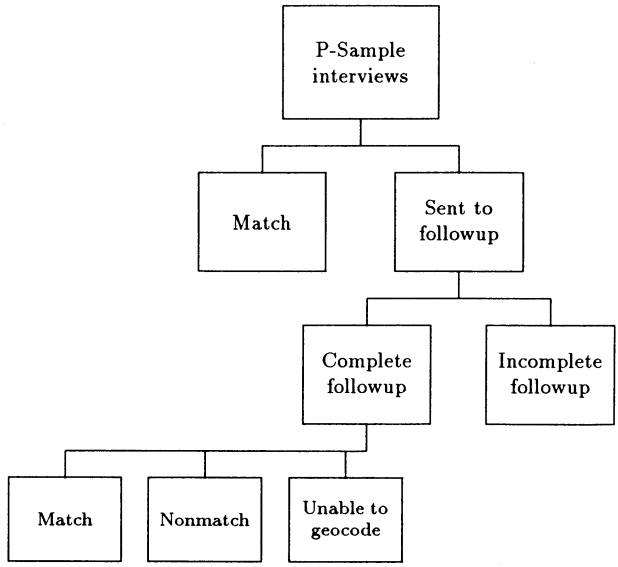

Fig. 1. Flow chart for P-sample interviews

Those persons whose Census Day addresses had been verified but who could still not be matched to the census were, at this point, considered to be resolved nonmatches, i.e., actual census omissions, and contributed to the estimates of undercount.

For a small number of cases, however, the additional information gathered on Census Day address through followup was still not specific enough to allow geocoding, the process by which a case was assigned a specific geographic location for a census record search. Virtually all of these cases were movers, having reported a Census Day address different from the CPS address. Geocoding could not be carried out for 3.3% of the April and 4.6% of the August cases whose followup interviews were completed. These ungeocodeable cases remained unresolved with respect to their final match statuses.

A flow chart of the P-sample interview cases, showing the possible outcomes of clerical match and followup, is given in Figure 1.

5. P-Sample Missing-Data Treatments

5.1. Noninterview weighting adjustment

The P-sample noninterviews were treated by a weighting adjustment. The noninterview households were removed from the sample, and their sample weights were redistributed across the interview households. This weighting adjustment was carried out within adjustment cells defined by geography and race. Since the majority of persons in the interview households were matched to the census, this weighting adjustment replaced the persons in noninterview households with persons who, for the most part, matched to the census. In other words, it effectively assumed a low census omission rate for noninterview persons, similar to the nonmatch rate for the sample as a whole.

5.2. Imputation of incomplete followups and ungeocodeables

For the incomplete followups and ungeocodeable cases, one treatment was an imputation procedure that filled in the missing

match statuses. An individual whose match status was unresolved was linked to a "donor," an individual who was similar with respect to other characteristics, but whose match status was resolved. The donor's match status was then imputed or assigned to the unresolved case.

Match status before followup played an important role in the linking of unresolved cases with donors. In the initial round of clerical matching, each case had been assigned one of the following prefollowup match codes:

M – Person was matched to the census.
N1 – Person did not match, but some other household members could be matched.
N2 – No household members were matched, but the household address was matched.
N3 – Neither the address nor any household members were matched.
PM – Probable match.

The imputation procedure proceeded in two steps, or waves. In the first wave, an outcome of "match," "nonmatch," or "ungeocodeable" was assigned to each incomplete followup case. This simulated the outcome of a followup interview which, had it been successfully completed, would have resulted in the case being resolved as a match, resolved as a nonmatch, or left unresolved due to geocoding problems. Each incomplete followup case was linked to a donor whose followup interview had been completed. This linking was based on demographic and geographic characteristics, and on the prefollowup match codes listed above.

In the second wave of imputation, an outcome of "match" or "nonmatch" was imputed to each ungeocodeable case, including those that had been imputed as "ungeocodeable" in the first wave. The linking of cases to donors in this second wave was based on demographic and geographic characteristics, and on mover status (whether or not the person had moved to the CPS address between Census Day and the CPS), but not on the prefollowup match code.

The results of the imputation for the April P-sample are shown in Figure 2. In the first wave, only a small fraction of the incomplete followups were imputed as "ungeocodeable." Of the rest, about half (52.0%) were imputed as "nonmatch," which roughly parallels the high nonmatch rate among the completed geocodeable followups (40.8%). In the second wave of imputation, "nonmatch" was imputed at rates much lower than in the first wave, both for the true ungeocodeables (29.5%) and for the cases imputed as "ungeocodeable" (13.3%). The lower nonmatch rate in the second wave is a consequence of the fact that prefollowup match code was not considered when cases were linked to donors in the second wave. The donor pool for the second wave included all the cases that had been matched without followup, whereas the donor pool for the first wave did not; hence, the second wave imputed nonmatches at a rate more similar to the low nonmatch rate of the entire sample.

An examination of Figure 1 reveals that only 3% of the interview cases from the April sample required imputation, but among all the cases finally treated as "nonmatch," 30% were unresolved cases imputed to nonmatch status. A similar pattern emerged in the August samples, where imputation of 3.3% of the cases contributed 25% of the total number of nonmatches. This result, which at first glance seems quite extreme, can be justified both theoretically and intuitively, and will be discussed in Section 7.

In an alternative missing-data treatment, the incomplete followups and ungeocodeable cases were included in the noninterview

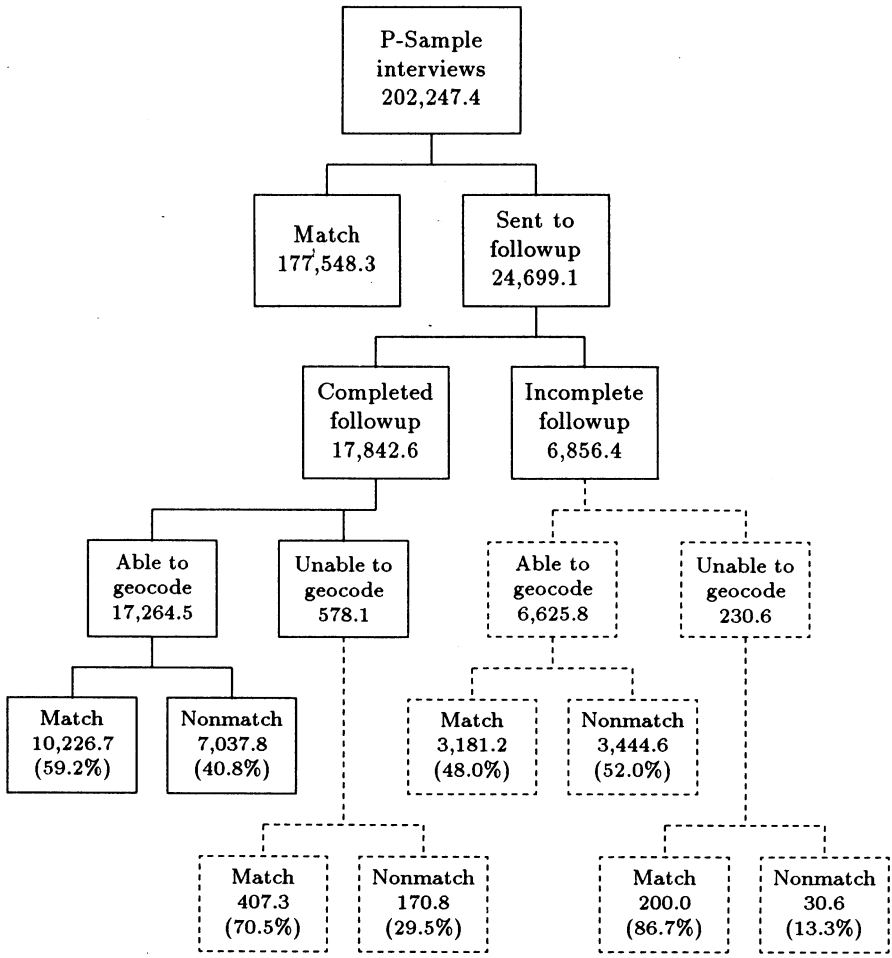

Fig. 2. Results of clerical match, followup, and imputation for the April P-sample interview cases. Numbers shown are estimates of the U.S. population in thousands, obtained by weighting individuals by their inverse probabilities of selection. Source: Fay (1988a), table 4.

weighing adjustment. The effect of this alternative treatment was to replace all the unresolved cases with resolved cases that matched at a higher rate, similar to that of the entire sample.

5.3. April noninterviews reconstructed from March and May

Because of the high rate of noninterviews, a "clean-up" effort was implemented in 1981 and 1982 to reconstruct the composition of some of the April noninterview households. The CPS is a panel survey with a rotation design in which a household is introduced into the survey for four consecutive months, excluded for eight months, and then reintroduced for a final four months; thus, each housing unit that was in the April CPS was also in the March CPS, the May CPS, or both. Many of the approximately 3,700 households that could not be interviewed in April had been interviewed in either March

or May, and about 2,700 could have their approximate composition reconstructed from those months.

These reconstructed households were subjected to clerical matching, and most persons in them were matched to the census. The remainder were designated for followup interviewing, but for most, no followups could be attempted for reasons of timing. In the end, 16% of these cases were unresolved with respect to match status; most were incomplete followups, but a few were ungeocodeable. These unresolved cases were included in the imputation described above for the unresolved interviews. Results of the imputation for the unresolved reconstructed April noninterviews are shown in Figure 3. A comparison of Figures 2 and 3 shows that nonmatches were imputed at similar rates in each case, which is not surprising since the donors were drawn from a common pool.

Including the data from these reconstructed households in the estimation, or simply removing all 3,700 April noninterview households by weighting adjustment, were two alternative missing-data treatments performed on the April P-sample. No corresponding effort was made to infer the household composition from adjacent months for the August P-sample.

5.4. *Problems associated with August movers*

In addition to the alternative missing-data treatments described above, another alternative involved the use of information about August movers. These were persons who, in the August interview, reported that they had lived at a different address on Census Day.

Some concern arose over the quality of geocoding for these cases. Among the movers considered to be resolved, the nonmatch rate was unusually high, in excess of 20%. It was thought that many of these nonmatches did not represent actual census omissions, but were the result of bad geocoding—they had been assigned to wrong geographical areas for matching. This would have happened if the information on Census Day address collected in the interview was incorrect, or if the geocoding process itself was unreliable. The problems associated with movers were thought to be more severe for August than for April, because the additional four months increased the rate of moving, and perhaps also increased the bias due to poor recollection of Census Day address.

In an alternative treatment, all movers were removed from the August P-sample. As a result, the reciprocal match rate N^P/M in the dual-system estimate (2) was estimated from nonmovers only, but was then applied to the entire population including movers. This treatment effectively assumed that the rates of census omission for movers and nonmovers were the same.

5.5. *Five P-sample sets*

Under the various alternative treatments for missing data, the Census Bureau prepared five separate P-sample datasets for dual-system estimation. Three of the sets used data from April, and the remaining two used data from August. The sets from April are:

Set 2 – Noninterview households reconstructed from March or May included; all other noninterviews removed by weighting adjustment; incomplete followups and ungeocodeables imputed.

Set 3 – Same as Set 2, but with the noninterview households reconstructed from March or May removed by weighting adjustment.

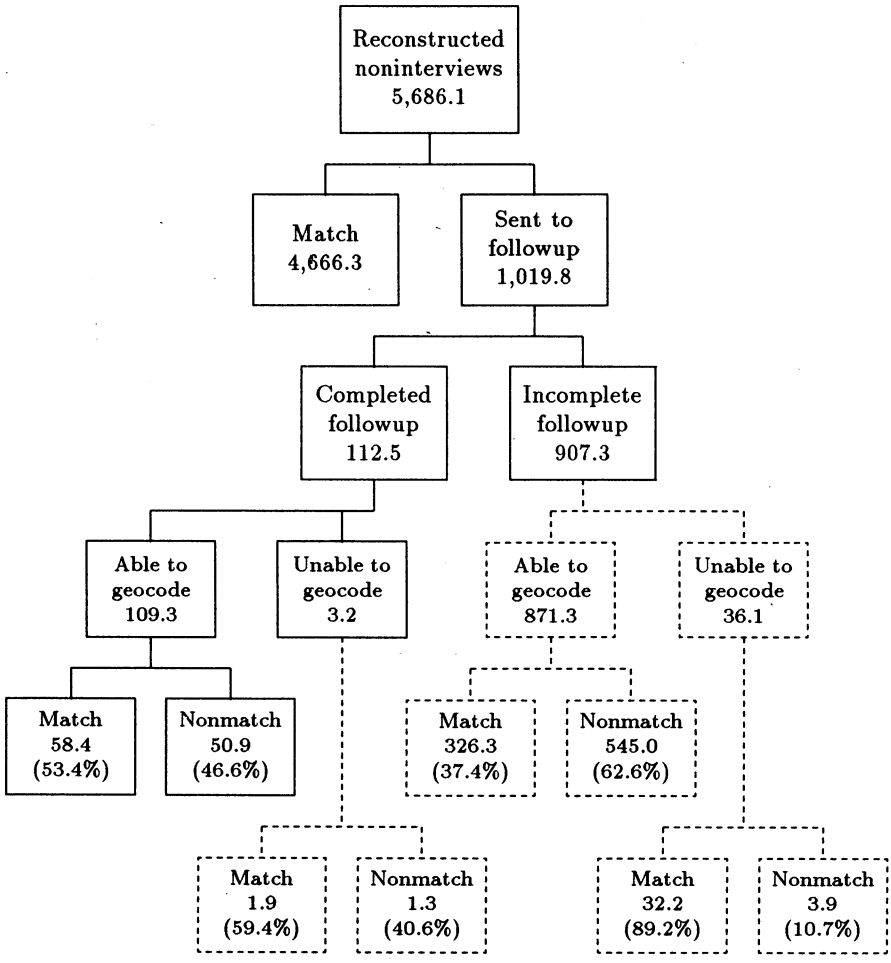

Fig. 3. Results of clerical match, followup, and imputation for the April P-sample noninterview households reconstructed from March and May. Numbers shown are estimates of the U.S. population in thousands, obtained by weighting individuals by their inverse probabilities of selection. Source: Fay (1988a), tables 4 and 8.

Set 14 – Same as Set 2, but with incomplete followups and ungeocodeables removed by weighting adjustment.

The sets from August are:

Set 5 – Noninterviews removed by weighting adjustment; incomplete followups and ungeocodeables imputed; movers included in the sample.

Set 10 – Same as Set 5, but with movers removed from the sample.

6. Criteria for Evaluating Missing-Data Treatments

6.1. *The use of probability models*

The classical methods for inference in sample surveys are based on the randomization used to draw the sample; see, for example, Cochran (1983). In this framework, an

estimator is evaluated according to its bias and variance, defined by the first two moments of its distribution over hypothetical repeated samples. When some survey data are missing, however, the "estimator" usually results from a series of rather complicated data manipulations, such as reweighting and imputation. Evaluating such an estimator is conceptually less straightforward, because in addition to error of random sampling, one needs to assess error arising from missing data and imperfect missing-data treatments.

A considerable literature has developed concerning missing-data techniques, both in sample surveys and in other statistical applications; see, for example, Little and Rubin (1987), and Rubin (1987). In this literature, joint probability models are specified for the data and for the nonresponse (or missingness) mechanism. Once a probability model is specified, the choice of missing-data techniques is guided by well-accepted statistical principles of efficient estimation and inference. The missing-data methods most commonly used in surveys, including imputation and reweighting techniques, can be motivated, at least approximately, as general purpose estimation techniques that remove nonresponse bias under certain probability models for the data and the nonresponse mechanism.

Some of the missing-data treatments employed in PEP were developed out of formal probability models for the data and nonresponse mechanism. Other treatments were thought of, not in terms of models, but simply as applications of commonly used missing-data procedures (Fay 1989). Each of the treatments, however, whether motivated by a probability model or not, would be appropriate under some model or class of models. Hence, we can compare the missing-data treatments by discussing the relative merits of the models that underly them. Without knowing the missing values themselves, or at least a sample of them, the nonresponse mechanism cannot be fully estimated by the data at hand. Such a discussion, therefore, will necessarily involve some subjectivity and conjecture. Nevertheless, we have a great deal of qualitative knowledge about the processes by which the PEP data were collected, and hence we can meaningfully discuss models in relation to situations that could have plausibly arisen in PEP.

6.2. *Ignorable and nonignorable models*

An important group of models for nonresponse is the class of ignorable models, as defined by Rubin (1976). In ignorable models, the probability that data are missing does not depend on the missing values, although it may possibly depend on observed values.

Virtually every missing-data procedure commonly used in survey practice is based on an assumption of ignorability. For example, a weighting adjustment that is carried out within adjustment cells assumes that data values and nonresponse are conditionally independent within these cells. Since the cells are formed on the basis of variables that are observed for all units, an assumption of ignorability has been made. Most imputation methods that link unresolved cases to donors based on characteristics observed for both also assumes ignorability. Many of the missing-data treatments used in PEP can be approximately motivated by models of ignorable nonresponse. It is crucial, therefore, to assess the suitability of these ignorable models.

6.3. *General ignorable procedures*

An important feature of ignorable models is that the pattern of response does not appear

	X	Y	R
units: 1	1	0	1
2	0	1	1
·	0	?	0
·	0	0	1
·	1	1	1
·	1	?	0
·	·	·	·
·	·	·	·
·	·	·	·
n	1	0	1

Fig. 4. Two binary variables, X and Y, where Y is subject to nonresponse

in the likelihood function used for likelihood-based inference about the parameters of the data model; see Little and Rubin (1987), Section 5.3. As a consequence, for a given data model, it is possible to design a missing-data procedure that will eliminate nonresponse bias under any model of ignorable nonresponse. We will call a missing-data procedure a general ignorable procedure if it has this property of eliminating nonresponse bias under any model of ignorable nonresponse.

A general ignorable procedure for sample surveys makes full use of information contained in survey variables that are observed for both respondents and nonrespondents. For illustration, consider a simple case of two variables, X and Y, where Y is missing for some units. For simplicity, we assume that both X and Y are binary, taking values 0 or 1. Let R be the response indicator, with $R = 1$ if Y is observed and $R = 0$ if Y is missing. A diagram of such a dataset is shown in Figure 4.

A general ignorable procedure for this dataset would use the information in X to adjust for missing data in Y. For example, a general ignorable weighting adjustment would form two adjustment cells, corresponding to $X = 0$ and $X = 1$. The nonrespondents would be removed from the sample and their sample weights redistributed among the respondents within these cells. A general ignorable imputation method would impute Y on the basis of X. For each nonrespondent, for example, a donor would be selected at random from all respondents who share the nonrespondent's value of X. If missingness of Y is related to X but not to Y, both the weighting adjustment and the imputation method can eliminate nonresponse bias.

6.4. Biases resulting from procedures that are less general

In the above example, a missing-data procedure that did not use the information in X would also remove bias under a special case of an ignorable model, although not under the most general one. A general ignorable procedure, however, would remove bias under either model. To illustrate this effect, we will compare the performance of two imputation procedures, one that uses information in X and one that does not.

To focus exclusively on nonresponse bias, suppose that the sample size n is very large, so that sampling variability is not an issue. Let the underlying joint distribution of X and Y in the sample (and in the population) be given by $p_{ij} = P(X = i,\ Y = j)$ for $i = 0,\ 1$ and $j = 0,\ 1$. Let r_{ij} denote the probability that Y is observed for a unit with $X = i$ and $Y = j$; that is, $r_{ij} = P(R = 1|X = i,\ Y = j)$ for $i = 0,\ 1$ and $j = 0,\ 1$. Under these frequencies, the observed distribution of data seen in the sample is easily calculated, and is shown in Table 2.

The values of r_{ij} determine the nonresponse mechanism. For now, we will consider two ignorable models:

1. Missing completely at random (MCAR): The probability of nonresponse depends neither on X nor Y;

Table 2. Frequencies observed in the sample

	$Y = 0$	$Y = 1$	Y missing
$X = 0$	$r_{00}p_{00}$	$r_{01}p_{01}$	$(1 - r_{00})p_{00} + (1 - r_{01})p_{01}$
$X = 1$	$r_{10}p_{10}$	$r_{11}p_{11}$	$(1 - r_{10})p_{10} + (1 - r_{11})p_{11}$

that is, $r_{00} = r_{01} = r_{10} = r_{11} = r$.

2. Missing at random (MAR): The probability of nonresponse may depend on X but not on Y; that is, $r_{00} = r_{01} = r_0$ and $r_{10} = r_{11} = r_1$.

Under MCAR, nonresponse is completely unrelated to both X and Y. Under MAR, nonresponse is related to X, and may also possibly be related to Y, but only through the variable X. Note that MCAR is a special case of MAR.

Consider first the imputation method that ignores X: Assign to each nonrespondent a value of Y from a respondent chosen completely at random. Suppose that we want to estimate the proportion of the population having $Y = 0$, which is $p_{00} + p_{10}$. If our estimate is the proportion of cases in the sample having $Y = 0$ after imputation, then straightforward algebra shows that the estimate will be, in expectation,

$$\frac{r_{00}p_{00} + r_{10}p_{10}}{r_{00}p_{00} + r_{10}p_{10} + r_{01}p_{01} + r_{11}p_{11}}. \quad (3)$$

Under MCAR, (3) reduces to $p_{00} + p_{10}$, and the procedure is unbiased. Under MAR, however, it is not generally equal to $p_{00} + p_{11}$. Now consider the imputation method that conditions on X to choose a donor for each nonrespondent; under this method, the estimate will be, in expectation,

$$p_{00}\frac{r_{00}(p_{00} + p_{01})}{r_{00}p_{00} + r_{01}p_{01}} + p_{10}\frac{r_{10}(p_{10} + p_{11})}{r_{10}p_{10} + r_{11}p_{11}} \quad (4)$$

which reduces to $p_{00} + p_{10}$ both under MCAR and MAR.

Another way to view these two models is to consider the three-way contingency table of complete data cross-classified by X, Y, and R. The MCAR model would include effects for X, Y, R, and $X \times Y$. The MAR model would include effects for X, Y, R, $X \times Y$, and $X \times R$. Both of these are natural models to investigate, and both can be estimated from the observed data; see, e.g., Fay (1986).

Note that the observed data can be used to distinguish MAR from MCAR. Under MCAR, the response fractions r_0 and r_1 should be equal, except for sampling variability. If they are significantly different, we can reject the MCAR hypothesis in favor of MAR. The observed data cannot, however, provide any evidence against the hypothesis of MAR.

6.5. The importance of conditioning on observed covariates for bias reduction

As illustrated above, missing-data procedures that condition on more observed covariates can eliminate bias under broader classes of nonresponse mechanisms. Conditioning on unnecessary covariates may, of course, increase the sampling variance of the final estimates, but it cannot introduce any bias. Suppose that, in the above example, MCAR were known to hold exactly; the conditional procedure would then be slightly less efficient than the unconditional procedure in finite samples, because it would fit

one unnecessary parameter (the $X \times R$ interaction) rather than setting it equal to zero. If MCAR were not known to hold, however, using the unconditional procedure rather than the conditional one could result in a substantial increase in bias. In large samples, therefore, where bias is the key issue, it is generally good policy to condition wherever possible in missing-data procedures. Even when an ignorable model does not exactly hold, conditioning on observed covariates will still tend to reduce bias, although it may not eliminate it entirely.

When the number of observed survey variables becomes very large, it may become difficult to condition on all of them, either in weighting adjustments or in imputation. Little (1986a) discusses techniques of data reduction that are useful in such cases.

In fairness, it should be mentioned that it is possible to construct a model under which an unconditional missing-data procedure would eliminate bias, but a conditional one would not. Such a model would be neither MCAR nor MAR, but would have nonignorable nonresponse. The model would have a large $X \times Y \times R$ interaction that *exactly* cancels out the $X \times Y$ interaction in the cross-section of the table for which $R = 0$. In other words, X and Y would have to be related to each other among the respondents, but exactly independent among the nonrespondents. Moreover, the marginal distribution of Y among the respondents and nonrespondents would need to be exactly the same. These two unusual constraints would make it unnatural to propose such a model a priori in most data analyses; one would almost need to believe in a malevolent state of nature that is trying to deceive the statistician.

In most cases, if a relationship between X and Y is observed among the respondents, it is quite natural to believe that a similar, although perhaps not identical, relationship would hold for the nonrespondents as well. Hence, lacking any special knowledge that such a relationship should not holds' for the nonrespondents, it is reasonable to use general ignorable procedures, because they make full use of the relationships among respondents to adjust for nonrespondents. On rare occasions, however, there may be a priori reasons to believe that a relationship observed among the respondents would not hold for nonrespondents. This situation did arise in PEP for the P-sample ungeocodeable cases, and will be discussed below.

7. Evaluating the P-Sample Treatments

7.1. Weighting adjustment for noninterviews

The P-sample weighting adjustment for noninterviews implicitly assumed that the noninterview households were like a random sample of all households within adjustment cells, or that the households', characteristics and their propensity to respond were conditionally independent within cells. This was a particular assumption of ignorability. If this condition held, then the weighting adjustment would have eliminated nonresponse bias; see, e.g., Little (1986b).

The assumption that the noninterview households were a random sample of all households within adjustment cells does not seem tenable. For example, experience has shown that nonresponding households tend to be smaller on average than responding ones. In the April P-sample, the interview households contained an average of 2.3 persons. Among the noninterview households whose composition was later reconstructed from March and May, however, the average household size was only 1.4 persons. Hence, the weighting adjustment

probably overstated the number of persons in the P-sample, because it replaced the noninterview households by interviewed households that tended to be larger. To the extent that census omission rates for persons in small households and large households differed, the weighting adjustment biased the estimates of undercount.

Other characteristics related to census omission were different for the interview and noninterview households as well. The combined effect of these differences probably made the noninterview adjustment too conservative, i.e., biasing the estimates of the undercount downward, although perhaps not dramatically. Evidence of this can be seen by comparing the data on April interviews, in Figure 2, with the data on April noninterviews reconstructed from March and May, in Figure 3. The reconstructed noninterviews that were resolved through followup had a somewhat higher nonmatch rate (46.6%) than the interviews resolved through followup (40.8%). Moreover, the imputations for the unresolved cases reveal higher imputed nonmatch rates for the reconstructed noninterviews than for the interviews; this shows that on the observed characteristics, the unresolved noninterviews tended to resemble nonmatch cases more that the unresolved interviews did. Hence, it may be reasonable to think that the mechanism relating census omission to nonresponse in the P-sample interviewing was nonignorable, and that the omission rates could have been systematically higher among the noninterviews than among the interviews. Households that were difficult to interview in the CPS were probably also difficult to enumerate in the census. To the extent that this was true, the noninterview adjustment would have been conservative.

7.2. The imputation of incomplete followups and ungeocodeables

The imputation for incomplete followups and ungeocodeables was based on an explicit model of nonignorable nonresponse. Fay and Cowan (1983) motivate the imputation method in terms of a recursive causal model for categorical data. A full description of their model is beyond the scope of this discussion; we will give only a brief intuitive justification for the method, emphasizing its nonignorable aspects.

As described in Section 5.2, the imputation procedure made an important distinction between the incomplete followups and ungeocodeables. For the incomplete followups, a status of "match," "nonmatch," or "ungeocodeable" was imputed in the first wave conditionally on prefollowup match code. For the ungeocodeables, including the cases imputed as "ungeocodeable" in the first wave, however, match status was imputed in the second wave without conditioning on prefollowup match code.

The fact that prefollowup match code was conditioned on for some cases but not for others caused this procedure to be nonignorable. Had it been conditioned on for all of the cases, this would have been essentially a general ignorable procedure. Had it been conditioned on for none of the cases, it would have also been an ignorable procedure, although a less general one. The practical effect of imputing the ungeocodeables unconditionally rather than conditionally is that matches were imputed to them at a higher rate, lowering the estimates of the undercount. The overall effect on coverage estimates could not have been very large, however, because as shown in Figure 2, ungeocodeables accounted for only about 10% of the unresolved cases.

The nonignorable model underlying the

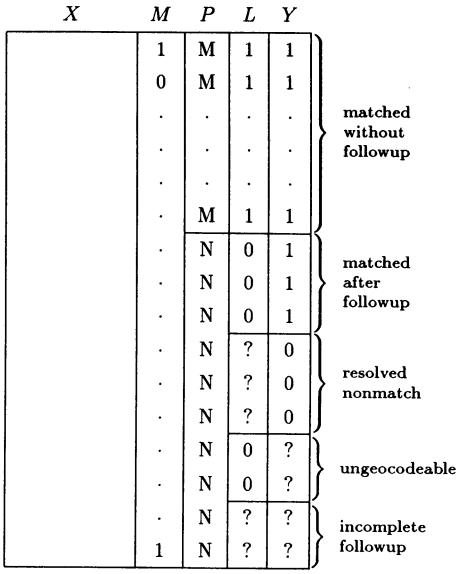

Fig. 5. Data observed for P-sample interview cases, including the latent variable L

imputation assumes an important relationship between prefollowup match code and final match status exists among the geocodeable cases, but that no relationship between these two variables exists among the ungeocodeables. In other words, this model states that a large interaction exists between prefollowup match code, final match status, and geocodeability, and that this interaction exactly cancels out any apparent relationship between prefollowup match code and final match status among the ungeocodeables. This model is analogous to the one described in Section 6.5 for which unconditional imputation would remove bias, but conditional imputation would not. To propose such a model without reason would be rather unusual, but in this case the model has a natural intuitive appeal.

The observed data from P-sample interviews is represented as a matrix in Figure 5, with rows representing individuals and columns representing survey variables. The variables are:

- X – Demographic and geographic characteristics.
- M – Mover status; 1 = P-sample mover, 0 = nonmover.
- P – Prefollowup match code; M = match, N = one of the four nonmatch codes listed in Section 5.2.
- Y – Final match status; 1 = match, 0 = resolved nonmatch.

An additional variable, denoted by L, also appears in Figure 5. As mentioned in Section 4.3, if a person was not matched in the prefollowup matching operation, it was for one of two principal reasons: either the person had been omitted from the census, or the correct Census Day address had not been searched. L is a latent variable that indicates whether the correct Census Day address had been searched before followup, with 1 indicating that the correct address had been searched, and 0 indicating that it had not (Fay 1989).

Although the value of L was not directly observed in the survey, it could be deduced for many of the cases. Those that were matched without followup obviously had $L = 1$. Many of those that were matched after followup probably had $L = 0$, because better information on Census Day address had been obtained through the followup interview to facilitate the match. For the cases that were ungeocodeable after followup, it is evident that the address information was poor before followup as well, so we can conclude that $L = 0$ for them. For cases that were resolved as nonmatches after followup, however, L is not really known, because whether or not better address information was obtained in the followup interview is not reflected in the dataset. L is also unknown for the incomplete followups,

because address information could not be verified through followup.

The variable L contains important information about the relationship between P and Y. Whenever $L = 1$, Y and P are related almost deterministically: $Y = 1$ when $P = \text{M}$, and $Y = 0$ for the majority of cases for which $P \neq \text{M}$. (When the matching clerks were searching the correct address, the fact that a case did not match would not necessarily have indicated census omission; e.g., some of the identifying characteristics of persons in the CPS or census may have been incorrect, missing, or discrepant. Hence, $L = 1$ and $P \neq \text{M}$ would not always imply that $Y = 0$, although it would indicate that $Y = 0$ with high probability.) When $L = 0$, however, P contains little or no information about Y, because the prefollowup matching attempt was futile; the clerks were not looking in the right place. Hence, we might reasonably believe that P and Y are independent in the cross-section of the contingency table for which $L = 0$, although a very strong, almost deterministic, relationship between them exists in the cross-section for which $L = 1$. Since the ungeocodeables have $L = 0$, imputing their values of Y unconditionally on P would remove nonresponse bias.

Because the relationship between P and Y is crucially influenced by L, an ignorable procedure that imputed missing values of both L and Y, taking full account of the joint relationships between L, Y, and the other variables, could have been appropriate. An examination of Figure 5, however, reveals that this was impossible for PEP, because certain portions of this joint distribution were inestimable. Specifically, the dataset contained no information about the conditional distribution of L given $Y = 0$, because L could not be deduced for any of the resolved nonmatches. Hence, had L been explicitly included in the imputation, there would have been no unique general ignorable procedure.

In subsequent post-enumeration surveys conducted by the Census Bureau, important design changes were introduced that deal with this issue more effectively. A number of new match codes were introduced, including codes for ungeocodeability at the prefollowup stage. The match rate among the prefollowup ungeocodeables that are resolved through followup now provides a basis for imputing the prefollowup ungeocodeables that are never resolved. Because of this greater level of detail in the recorded data, the concerns over geocodeability that were handled by a nonignorable method in 1980 can now be handled by straightforward ignorable methods.

7.3. *The use of reconstructed April noninterviews*

As described in Section 5.3, about 73% of the April noninterviews could have their household composition reconstructed from March and May, and to include these households or weight them out of the sample constituted two alternative treatments. P-sample Set 3, which weighted them out of the sample, lowered the estimated undercount rates by 0.1% overall and 0.4% for blacks relative to Set 2, which did not weight them out, a movement of about one-half of a sampling standard error. The reason for this is apparent from Figures 2 and 3. Although the nonmatch rates among the reconstructed noninterviews were somewhat higher than for the interviews, the differences are not so dramatic.

What issue is being addressed by these two alternative treatments is not entirely clear. A comparison of Sets 2 and 3 is very interesting from a standpoint of survey design, because it suggests that the extra effort required to reconstruct the data for

April noninterviews had only a minimal effect on the quality of the final estimates. The comparison also has important implications for the weighting adjustment used for the remaining noninterviews, because it suggests that the noninterviews (at least the ones that could be reconstructed) were not dramatically different from the interviews with respect to census undercount; hence, the noninterview adjustment was probably not unreasonable. A comparison of Sets 2 and 3 also suggests what the results of PEP might have been if data had been available for all noninterviews. Since the April noninterview households that were interviewed in March or May had higher rates of undercount than the April interview households, we might suspect that the April noninterviews that were not interviewed in March or May may have been undercounted at even higher rates, because they represent persons for whom data collection is even more difficult.

To the question of whether or not these data ought to be included in the best estimates of coverage, however, the answer seems obvious. Noninterview households are an important part of the universe to which the results of this survey were meant to generalize. Omitting the reconstructed data, unless they are seriously biased by measurement error, could hardly be expected to improve the estimates. To the knowledge of this author, no one has ever presented a strong case for why the data for these reconstructed households should be omitted. To the extent that the noninterview population was better represented by the reconstructed noninterviews than by the interviews, omitting these cases could have only increased nonresponse bias, and also reduced the precision of the final estimates. For purposes of inference, then, Set 2 seems preferable to Set 3.

7.4. The use of August movers

Section 5.4 described the concern about the movers in the August sample: Many of the cases called "resolved nonmatches" might have represented bad geocoding rather than census omission. P-sample Set 5 included August movers, but Set 10 eliminated them entirely. Omitting movers lowers the estimated undercount rates by 1.4% overall and 1.7% for blacks. Interestingly, although omitting movers made a rather large difference in the estimated level of undercount, it had a relatively smaller effect on estimated differences in undercounting among the important subpopulations.

Many of the comments regarding the use of reconstructed April noninterviews apply to August movers as well. The population of movers is an important part of the universe to which we wish to generalize, and hence it would be desirable to include information on movers if at all possible. In the case of movers, however, substantial measurement error in the data is a distinct possibility.

The practice of omitting all movers from the sample seems to be an overreaction to the real issue of concern. The reliability of final match status was not in question for all movers, but only for the movers that were not matched to the census and were not classified as ungeocodeable—i.e., the resolved mover nonmatches. A sensitivity analysis that examined the effect of plausible alternative assumptions about this much smaller group of cases would have given a more accurate picture of the real level of uncertainty than simply omitting all movers.

We may suppose that an unknown proportion, say α, of the resolved mover nonmatches represent genuinely identified census omissions, with the remaining proportion $1 - \alpha$ representing bad geocoding. Since any mover who had been geocoded

correctly and who had been omitted from the census would have been classified as a resolved nonmatch, we would have to presume that α is greater than zero. Among the remaining proportion $1 - \alpha$ of cases, not all could have been census inclusions; since the correct Census Day address had never been searched for these cases, the prefollowup status of nonmatch is, at worst, meaningless, and it would be conservative to assume that these matched to the census at a rate similar to the sample as a whole. An analysis carried out under these assumptions, varying α over a plausible range of values, would have more accurately reflected the concern about data quality for movers. Under such an analysis, the undercount estimates would have probably changed far less than the discrepancy between Sets 5 and 10.

7.5. *The weighting adjustment of Set 14*

The only missing-data treatment yet to be discussed is the weighting adjustment of P-sample Set 14. In this treatment, all unresolved cases, including noninterviews, incomplete followups, and ungeocodeables, were treated as noninterviews and weighted out of the sample. The effect of this alternative was to lower the estimated national undercount rate by more than 2% overall, making it into a net national overcount. This single treatment, more than any other, was a primary cause for concern in the interpretation of the results of PEP; if Set 14 represented a plausible alternative, then we can only conclude that PEP's margin of error was too great to allow any conclusions about the magnitude of the undercount in the 1980 census.

The critical feature of the Set 14 procedure that caused its undercount estimates to be so much lower than Set 2's was that it was unconditional; it did not make use of covariate information available for the unresolved interview cases. Specifically, it did not condition on the most crucial indicator of final match status, the prefollowup match code. By combining data from Figures 2 and 3, we can construct a frequency table of prefollowup match code by final match status observed in the April P-sample, as shown in Table 3. This simple representation of the April data has the same structure as the illustrative example of Sections 6.3–6.5.

An unconditional missing-data procedure for this dataset would infer the missing values of Y from the marginal distribution of Y among the resolved cases. Since 3.5% of the resolved cases have $Y = 0$, an unconditional procedure would assume a 3.5% nonmatch rate among the unresolved cases as well, leading to an estimated gross undercount of 3.5% for the entire sample. A general ignorable procedure, on the other hand, would condition on P to infer the missing values of Y. Since all unresolved cases have $P = \text{N}$, and since the resolved cases with $P = \text{N}$ are 41% nonmatches, the conditional procedure would assume a 41% nonmatch rate among the unresolved cases as well, leading to an estimated gross undercount of 5.0% for the entire sample.

The Set 14 weighting adjustment was very much like the unconditional procedure described above, because it did not condition on prefollowup match status; the adjustment cells were based only on coarse geography and race. The Set 2 imputation fell somewhere between the unconditional and conditional procedures, because it conditioned on prefollowup match code for geocodeables but not for ungeocodeables; since some 90% of the unresolved cases were deemed to be geocodeable, however, it was much closer to the conditional procedure.

The Set 14 weighting adjustment did not arise from an explicit probability model for

Table 3. Prefollowup match code P(M = match, N = nonmatch) by final match status Y(1 = match, 0 = nonmatch) as observed in the April P-sample, including interviews and reconstructed noninterviews. Numbers shown are estimates of the U.S. population in thousands, weighted by inverse probability of selection. Source: Fay (1988a) tables 4 and 8.

	$Y = 0$	$Y = 1$	Y missing
$P = \text{M}$	0	182,214.6	0
$P = \text{N}$	7,088.7	10,285.1	8,345.0

the data or the nonresponse mechanism, but was regarded merely as an application of the common survey practice of reweighting for unit nonresponse (Fay 1989). Nevertheless, it is useful to consider under what models this procedure would have been appropriate. It would have been appropriate under the model of MCAR, but by simply examining the observed data we can absolutely reject the MCAR hypothesis; all of the unresolved cases had $P = \text{N}$. Hence, the only conceivable model for Set 14 is the strongly nonignorable model described in Section 6.5, in which P and Y are dependent among the resolved cases but independent among the unresolved cases, and in which the marginal distribution of Y is the same in both groups.

In order to believe that the Set 14 estimates were not badly biased, then, we would need to believe that P was completely uninformative for predicting Y among all the unresolved cases, even though we can see that it is strongly informative for predicting Y among the resolved cases. This is a believable hypothesis for the ungeocodeables, for whom the correct Census Day address had never been searched before followup. It may also be a believable hypothesis for some other groups of unresolved cases as well. If some of the P-sample interviewers had fabricated data in the original interviewing, these fabricated persons would have shown up as prefollowup nonmatches, and this status of nonmatch would have little relationship to the actual match status of the real persons who had lived there. Fay (1989) suggests that this hypothesis could also apply to "hypermovers," persons who moved into a P-sample housing unit between Census Day and the April interview, were present for the April interview, and then moved out before followup.

For Set 14 to have been appropriate, however, we must not only believe that P and Y were unrelated among the groups of unresolved cases mentioned above, but also for every other unresolved case in the P-sample as well. This is hardly plausible. We know that there are many substantial groups of unresolved cases for which prefollowup match status must be informative; for example, any case that represented an actual census omission at the P-sample address, but became unresolved due to followup incompleteness, would have had an informative prefollowup match code. To believe the independence hypothesis for every single unresolved case is quite extreme. To this author's knowledge, no one has ever constructed a plausible scenario under which the model underlying P-sample Set 14 would have been true.

8. Conclusions

To summarize, the following five points have been made regarding nonresponse bias in the P-sample missing-data treatments:

1. The noninterview weighting adjustment employed in all five P-sample sets was probably not unreasonable, in light of the limited evidence about April noninterviews whose data was reconstructed from adjacent months. The differences in nonmatch rates among the resolved interviews and noninterviews suggest that it could have been somewhat conservative, causing the undercount estimates to be too low.
2. The imputation procedure for incomplete followups and ungeocodeables used in every set, except Set 14, was based on a nonignorable model that properly distinguished the geocodeables from the ungeocodeables. Matches for the ungeocodeables were imputed at a much higher rate, because the prefollowup code of nonmatch for these cases was not informative.
3. Omitting the reconstructed April noninterviews was unlikely to improve the estimates, and hence Set 2 seems preferable to Set 3.
4. Omitting the August movers was probably an overreaction to a legitimate concern. A more principled sensitivity analysis might have changed the estimates much less than the difference between Sets 5 and 10.
5. The weighting adjustment of Set 14 is based on the implausible model that prefollowup match code was uninformative for every unresolved case; this probably introduced substantial downward bias into the estimates of undercount.

Taken together, these five points would suggest that among the five P-sample datasets, Sets 2 and 5 contain the least nonresponse bias. With the exception of the first point, however, nothing has been said about the magnitude or direction of potential biases that are common to all of these sets. The nonignorable aspect of the models underlying the Set 14 weighting adjustment and, to a lesser extent, the imputation, tended to pull the undercount estimates downward, below what would be expected from a general ignorable procedure. Other plausible nonignorable models, however, could have easily raised the undercount estimates above those of a general ignorable procedure.

The purpose of the P-sample operations was to identify the sampled persons as counted or omitted in the census. The population of persons omitted from the census is, from the standpoint of census methodology, a group for which data collection is notoriously difficult; they may be uncooperative, unable to speak English, unaware of the data collection efforts, highly mobile, rarely at home, or may have any of a host of other qualities that make them difficult to track or hard to reach. It is plausible to believe that many of the same factors that caused persons to be omitted from the census could also have led them to be unresolved in the P-sample.

The P-sample measured a limited number of characteristics other than match or nonmatch to the census, including age, sex, race, mover status, etc. As more information on these recorded characteristics was included in the missing-data treatments, the estimated rates of undercount always tended to rise. This shows that the unresolved persons tended to resemble persons omitted from the census in the distributions of these recorded characteristics. One could imagine a host of other characteristics, ones that are less easily measured and were not recorded in the P-sample, which, had they been conditioned upon the missing-data treatments, would have raised the undercount estimates even further. Had these variables been

recorded in the dataset, a general ignorable procedure could have eliminated any nonresponse bias related to them. Since these variables were not recorded, however, the only way that nonresponse bias related to them could have been eliminated would have been through a nonignorable procedure, one that would have raised the gross undercount estimates above any of the five P-sample sets.

It should be noted, however, that the types of nonignorable procedures mentioned above, which would have tended to raise the estimates of gross undercount in the P-sample, would also have tended to raise the estimates of gross overcount in the E-sample. The E-sample, which measured erroneous enumerations, was an effort to collect data on another notoriously difficult group, including fictitious persons, out-of-scope persons, etc. As more recorded variables were conditioned upon in the E-sample missing-data treatments, the estimates of gross overcount also tended to rise. It is possible that nonignorable procedures that would have eliminated nonresponse bias in the E-sample would also have given estimates of gross overcount above any of the three E-sample sets. The combined effect that these nonignorable missing-data treatments would have had on the PEP estimates of net undercount is not entirely clear.

9. References

ASA Technical Panel on the Census Undercount (1984). Report, American Statistician, 38, 252-256.

ASA Technical Panel on the Census Undercount (1985). Correction, American Statistician, 39, 241.

Bailar, B.A. (1985). Comment on Estimating Population in a Census Year: 1980 and Beyond, by E.P. Ericksen and J.B. Kadane. Journal of the American Statistical Association, 80, 109-114.

Cochran, W.G. (1983). Sampling Techniques. New York: Wiley.

Ericksen, E.P. (1983). Affidavit Submitted to U.S. District Court, Southern District of New York. In *Cuomo v. Baldridge*, 80 Civ. 4550 (JES).

Ericksen, E.P. (1986). Comment on Regression Models for Adjusting the 1980 Census, by D.A. Freedman and W.C. Navidi. Statistical Science, 1, 18–21.

Ericksen, E.P. and Kadane, J.B. (1985). Estimating the Population in a Census Year: 1980 and Beyond (with discussion). Journal of the American Statistical Association, 80, 98–131.

Ericksen, E.P., Kadane, J.B., and Tukey, J.W. (1989). Adjusting the 1980 Census of Population and Housing. Journal of the American Statistical Association, 84, 927–944.

Fay, R.E. (1985). Comment on Estimating Population in a Census Year: 1980 and Beyond, by E.P. Ericksen and J.B. Kadane. Journal of the American Statistical Association, 80, 114–116.

Fay, R.E. (1986). Causal Models for Patterns of Nonresponse. Journal of the American Statistical Association, 81, 354–365.

Fay, R.E. (1988a). Evaluation of the Census Coverage from the 1980 Post Enumeration Program (PEP): Missing Data in the P-Sample. Preliminary Results Memorandum No. 123, U.S. Bureau of Census, Washington, DC.

Fay, R.E. (1988b). Evaluation of the Census Coverage from the 1980 Post Enumeration Program (PEP): Missing Data in the E-Sample. Preliminary Results Memorandum No. 126, U.S. Bureau of the Census, Washington, DC.

Fay, R.E. (1989). Comments on a Report by Joseph L. Schafer, unpublished memorandum dated 2/15/89, U.S. Bureau of the

Census, Washington, DC.

Fay, R.E. and Cowan, C.D. (1983). Missing Data Problems in Coverage Evaluation Studies. Proceedings of the Section on Survey Research Methods, American Statistical Association, 158–163.

Fay, R.E., Passell J.S., and Robinson, J.G. (1988). The Coverage of Population in the 1980 Census. 1980 Census of Population and Housing, PHC80-E4, U.S. Department of Commerce.

Freedman, D.A. and Navidi, W.C. (1986). Regression Models for Adjusting the 1980 Census (with discussion). Statistical Science, 1, 1–39.

Kadane, J.B. (1986). Comment on Regression Models for Adjusting the 1980 Census, by D.A. Freedman and W.C. Navidi. Statistical Science, 1, 12–17.

Little, R.J.A. (1986a). Missing Data in Census Bureau Surveys. Proceedings of the Second Annual Research Conference, U.S. Bureau of the Census, 442–454.

Little, R.J.A. (1986b). Survey Nonresponse Adjustments for Estimates of Means. International Statistical Review, 54, 139–157.

Little, R.J.A. and Rubin, D.B. (1987). Statistical Analysis with Missing Data. New York: Wiley.

Madansky, A. (1986). Comment on Regression Models for Adjusting the 1980 Census, by D.A. Freedman and W.C. Navidi. Statistical Science, 1, 28–30.

Rubin, D.B. (1976). Inference and Missing Data (with discussion). Biometrika, 63, 581–592.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

Schafer, J.L. (1988). Missing-Data Procedures in the 1980 Post-Enumeration Program: Why P-Sample Set 14 Should Not Be Trusted. Unpublished technical report, Department of Statistics, Harvard University, Cambridge, MA.

Wachter, K.W. (1983). Affidavit Submitted to U.S. District Court, Southern District of New York. In *Cuomo v. Baldridge*, 80 Civ. 4550 (JES).

Wolter, K.M. (1986a). Some Coverage Error Models for Census Data. Journal of the American Statistical Association, 81, 338–346.

Wolter, K.M. (1986b). Comment on Regression Models for Adjusting the 1980 Census, by D.A. Freedman and W.C. Navidi. Statistical Science, 1, 24–28.

Received January 1991
Revised November 1991