Missing Data in the 1988 Dress Rehearsal Post-Enumeration Survey

Gregg J. Diffendal, Bureau of the Census*, Washington, DC 20233

1. Introduction

This paper describes the levels of missing data and the methods for dealing with missing data in the 1988 Dress Rehearsal Post-Enumeration Survey (PES). The 1988 PES tested the procedures and methodologies that will be used to estimate census coverage in the 1990 Decennial Census. The 1988 Dress Rehearsal consisted of three test sites: St. Louis, East Central Missouri, and Eastern Washington.

Schenker (1988) gives a description of the missing data procedures in the 1986 PES in Central Los Angles County and the missing data rates. Similar procedures were used in the 1988 PES. Modifications were made in the noninterview adjustments and in other areas because of changes in the procedures conducted in the field, changes in the cases sent to followup and changes in the definitions in the poststrata. Childers and Hogan (1989) contains a complete description of the 1988 PES. The 1988 PES included certain segments of the group quarters population. This segment of the population were not included in the 1986 PES, specifically the non-institutional, nonmilitary group quarters population.

The missing data problems in the 1988 PES arise from noninterviews in the P(population) sample, missing data characteristics on the housing and person characteristics in both the P and E(enumeration) samples, missing match status for the P sample and missing enumeration status in the E sample. Each of theses missing data problems are discussed in the following sections.

2. Noninterviews in the P Sample

Weighting adjustments were used for all P-sample noninterviews, P-sample interviews recorded as last resort, whole household P-sample duplicate enumeration and whole household P-sample fictitious enumeration. All of these cases were treated as noninterviews. Table 1 shows that of the housing units judged to be occupied, 2.2% in St. Louis, 0.4% in East Central Missouri and 0.2% in Eastern Washington were considered as noninterviews.

The block sample design was used in 1988, as was used in 1986 and will be used in 1990, which allows for a means for handling P-sample noninterviews. In addition, the housing unit structure question was interviewer filled rather than respondent filled as in 1986. Therefore the noninterview weighting adjustment assumes that the distribution of the noninterviewed household is the same as the interviewed households within the block and the type of structure. To prevent any household from getting too large a weight from the noninterview adjustment, the limit was set so that an interviewed household could only represent two other noninterviewed household. If the limits were exceeded, then the weight was distributed over a larger number of interviewed households. This prevents any small number of housing units having a weight being much larger than other housing units in the same stratum.

3. Missing Characteristics in the P and E Samples

Missing characteristics occur even when the households are interviewed. Respondent refusals, miscoding, or interviewer error are some of the reasons for missing characteristics.

The test sites were poststratified by age (1 = 0-9, 2 = 10-19, 3 = 20-29, 4 = 30-44, 5 = 45-64, 6 = 65+) race (1=white nonhispanic, 2=all other race/ethnic groups), sex (1=male, 2=female) and tenure (1=owned, 2=rented). Tenure was only used in the poststratification in the St. Louis test site. These characteristics needed to be imputed before the dual-system estimates could be calculated by these variables. Table 2 contains the missing data counts for these variables in the P and E samples for the three test sites.

For the P sample, the largest missing characteristic was race with 1.2 percent missing in St. Louis and Eastern Washington. The E sample had more missing characteristics than the P sample for all of the variables used in the dual-system estimator. The largest missing characteristic was race with 3.8% in Columbia.

Missing characteristics were imputed using the hot-deck method similar to the procedure discussed in Schenker (1988).

Tenure, structure, and race were imputed using the most recently observed value, assuming that these variables are related geographically. Sex was imputed as the opposite sex if the spouse's sex was observed. Otherwise it was imputed at random from the distribution of all observed cases.

Age was randomly imputed from the distribution of all observed cases with the same type of household (single person/multiperson), marital status, relationship to head of household, and age of head of household.

Table 1 Household Interviews for the Three Test Sites

	St. Louis	E.C. Missouri	Eastern Washington
TOTAL Completed Interviews	4904	3396	919
Interview with Household	4660	3333	881
Proxy Interview	244 (4.9%)	63 (1.8%)	38 (4.1%)
Noninterview Total	108 (2.2%)	15 (0.4%)	2 (0.2%)
Noninterview	65	10	1
Last Resort and other	43	5	1
Total Occupied HU	5012	3411	921
Vacant	806	657	192
Total HU	5818	4068	1113

Table 2 Missing Characteristics for the Three Test Sites

P sample

	St. Louis	E. C. Missouri	Eastern Washington
	(12507)	(8225)	(2392)
Tenure	156(1.2)	71(0.9)	7(0.3)
Sex	59(0.5)	15(0.2)	1(0.0)
Age	61(0.5)	11(0.1)	8(0.3)
Race	153(1.2)	53(0.6)	29(1.2)

E sample

	St. Louis	E.C. Missouri	Eastern Washington
	(13581)	(9125)	(2628)
Tenure	279(2.1)	105(1.2)	45(1.7)
Sex	82(0.6)	30(0.3)	11(0.4)
Age	387(2.8)	140(1.5)	42(1.6)
Race	269(2.0)	345(3.8)	92(3.5)

Table 3 Missing Data in the P-sample Match Status

	St. Louis	E.C. Missouri	Eastern Washington
Match Status Unresolved	230(1.8)	230(2.8)	44(1.8)
Imputed Matched	68.5(29.8)	145.6(63.3)	19.9(45.2)

4. Missing P-Sample Match Status

For missing match statuses, a logistic regression model was used to predict the probability of being matched. The probability of being matched is used in the estimation phase, rather than a nonmatch (zero) or matched (one) which is used for the resolved cases.

Separate logistic regression equations were fit for each of the three test sites. All resolved cases were used as data points to estimate the regression coefficients. The variables used were: proxy (proxy interview, nonproxy interview), tenure (owner, renter), structure (single unit, not single unit), sampling stratum, sex, age (0-9, 10-19, 20-29, 30-44, 45-65, 65+), race, mover, and followup indicator (not followed up, followed up with housing unit match, other). The estimated parameters are shown in table 5. For a small number of cases, the mover status and followed up status were unknown. In these cases the logistic regression model was refit to the data with those two variables removed.

The match status was unresolved for 230(1.8%) of the 12507 persons in St. Louis, for 230(2.8%) of the 8225 persons in East Central Missouri and for 44 (1.8%) of the 2392 persons in Eastern Washington. The sum of the imputed match probabilities for St. Louis was 68.5(29.8%), for East Central Missouri was 145.6(63.3%) and for Eastern Washington was 19.9(45.2%). For the resolved cases the match rates were 86.8, 93.5, and 90.0 for the three test sites respectively. The large percent differences in the imputed match probabilities are mainly due to low nonmatch rates for the followup interview. In addition, most unresolved cases from followup were movers since all nonmatched movers were sent to followup.

An unresolved mover status occurs when it is not possible to determine which address the person should have been counted. This occurs when there is uncertainty in the interpretation of census residency rules. Only 25 persons in all three test sites had an unresolved mover status. All had imputed match statuses.

5. Missing E-Sample Enumeration Status

A census enumeration is considered to be correct if three criterion are met. First, the address where the person was enumerated has to be the correct address by census residency rules. For example, a college student enumerated at his parents house is labeled an erroneous enumeration even if the college student was not enumerated at his college address. A related aspect is that the person exists (not curbstoned). Second, the address has to be placed in the correct census geography. A search area is defined around a sample area so that minor miscoding in census geography are not considered as errors. Third, the census person has to be counted only once in the search area. If a person is duplicated within the search area, one person is erroneously enumerated and the other is correct, at least by the third criterion. However, rather than trying to make this determination, the sampled person is assigned as having a probability of one-half of being correctly enumerated.

Each of the three criterion must be met in order for the sampled person to be correctly enumerated. Missing data arises from the first and second criterion. Duplication, the third criterion, is not imputed, although the within block duplicates are used in predicting missing enumeration statuses.

A logistic regression model was used to predict missing enumeration statuses for the first criterion (within block duplicates were coded as observed and as erroneously enumerated). For census geography, only 4 persons were erroneously geocoded in St. Louis, 23 in East Central Missouri and 0 in Eastern Washington. Therefore, the estimated probability of being correctly geocoded was used rather than any logistic regression equation.

The estimated probability of being correctly enumerated was defined as

$$P(CE)=P(ES)x\ P(CG)x\ \frac{1}{1+DUP}$$

P(ES) is the observed or predicted probability of the enumeration status.
P(GC) is the observed or predicted probability of being correctly geocoded.
DUP is the number of duplicates in the surrounding blocks.
If P(ES) or P(GC) is equal to zero, the case in coded as an erroneous enumeration.

Table 4 Missing Data for E-Sample Status

	St. Louis	E.C. Missouri	E. Washington
Missing Enumeration Status	311	81	54
Imputed as Correct Enumeration	272.4 (87.6)	73 (90.1)	51.3 (95)
Missing Geocodes	30	25	0
Imputed Correctly Geocodes	29.7 (99)	24.5 (98)	0
Census Duplicates in Surrounding Blocks	44	92	12

In one block cluster in East Central Missouri and one in Eastern Washington, a large number of housing units were recorded in the census with no housing units recorded in the P sample. To reduce the followup workload, these two block clusters were subsampled before they were sent to followup. This occurrence had not been foreseen. Therefore, rather than weighting the observed cases in the estimation phase, all of these cases were imputed. Unfortunately, a processing error for the block cluster in East Central Missouri occured and none of these cases were sent to followup. The cases that should have gone to followup were imputed and the average imputed value for the block cluster was used to impute the rest of the block cluster. In East Central Missouri the probability of being correctly enumerated was estimated to be 0.82 and in Eastern Washington the probability was estimated to be 0.93.

Plans for 1990 are to use a weighting adjustment for block clusters with no P-sample housing units and a larger number of census housing units that are subsampled for followup.

Table 5 P-Sample Logistic Regression Equations

Variables	St. Louis	East Central Missouri	Eastern Washington
intercept	2.75	-0.23	-2.70
Proxy	0.98	0.59	0.23
Tenure	0.13	0.48	0.60
structure	0.40	0.33	-0.11
STRATUM 1	-0.40	1.46	2.51
STRATUM 2	-0.49	1.47	3.43
STRATUM 3	-	1.4	2.50
SEX	-0.18	-0.10	-0.08
AGE 0-9	-1.06	-1.01	-0.53
AGE 10-19	-0.93	-1.13	-0.08
AGE 20-29	-1.12	-1.15	-1.02
AGE 30-44	-0.81	-0.69	-0.42
AGE 45-64	-0.36	-0.28	-0.37
RACE	0.67	0.21	0.77
MOVER	-0.97	0.56	-0.28
Followup Housing Unit Match	-3.98	-2.17	-0.45
NOT followed up	0.24	1.68	2.18

6. Alternate Treatment of Missing Data

This section describes three alternate treatments of missing data in the 1988 PES. The first alternate treatment treats all proxy interviews as noninterviews. Proxy responses may have higher nonmatch rates due to incomplete or incorrect information on census day address. An alternate explanation is that these households are hard to contact and the persons in these households are more likely to be missed in the census.

The second alternate treatment treats all movers who lived out of the test site as missing match statuses. These persons were treated as out-of-scope in the PES. In 1990 all of these persons would be searched at their census day address and will be part of the PES and many would be resolved. The alternate treatment imputes a match status for all out-of-scope movers, assuming all are movers with an unresolved match status.

The third alternate treatment uses the original coding for movers. Because of difficulties in correctly coding movers, all nonmatched movers were rematched for the final PES estimates. This treatment uses the coding before this rematch. The differences between estimates of the 88 PES and the original mover coding shows some of the uncertainty in handling movers.

Table 6 shows the dual-system estimates for the 1988 PES and for the alternate treatments of missing data labeled proxy, movers, and original mover coding. For the proxy treatment, larger undercounts are estimated for the all other renter strata and both white nonhispanic strata in East Central Missouri. Lower undercount are estimated for the other strata. The differences in the estimates from the 1988 PES and the proxy treatment is fairly small. The higher undercount estimates are surprising, usually the reverse is expected. This implies that proxy respondents had higher match rates that other persons within the block cluster and type of structure.

For the mover treatment, larger undercounts are estimated for all strata. Since movers had lower match rates than nonmovers, this results is not surprising. For St. Louis and East Washington the effects of missing data from the mover treatment is small. For East Central Missouri the effects are much larger, some estimating an undercount rate of about twice as large as the 1988 PES estimates.

For the original mover treatment, the estimated undercounts do not differ significantly from the other alternate treatments. A higher undercount is estimated for white nonhispanic in TAR areas in East Central Missouri and for not in list/enumerate areas in Eastern Washington. The results show some uncertainties in the matching for movers, typically the hardest group to match due to geocoding and more difficult access to census materials.

These three alternate treatments do not reflect the full amount of uncertainty in the 1988 PES due to missing data. For example, other alternate treatments of missing data could use information from the followup interview. Proxy responses in the P-sample and E-sample followup could be assumed to be missing and imputed from the household respondents. Followup information that is not contained in the match codes was not used in the estimates presented here.

Table 6 Dual System Estimates for the 1988 PES and Three Alternate Treatments of Missing Data
(Undercount Rate in Parentheses)

	1988 PES	Proxy	Mover	Original Mover Coding
St. Louis				
WNH renter	94359(5.9)	94066(5.6)	94900(6.4)	93990(5.5)
WNH owner	116100(-1.2)	116075(-1.2)	116296(-1.0)	116232(-1.1)
All other renter	127776(11.5)	127805(11.5)	128572(12.0)	128268(11.8)
All other owner	87760(8.4)	87630(8.3)	87949(8.6)	87788(8.5)
East Central Missouri				
WNH TAR	67542(8.5)	68665(10.0)	68708(10.1)	69101(10.6)
WNH not TAR	356405(4.3)	359158(5.1)	369753(7.8)	357253(4.6)
All Other	28525(10.8)	28482(10.7)	29885(14.9)	28895(12.0)
Eastern Washington				
L/E	139655(7.3)	138863(6.8)	140403(7.8)	139924(7.5)
Not L/E	135958(6.3)	134689(5.5)	136102(6.5)	137578(7.5)

WNH - white nonhispanic
TAR - Tape address registers (urban areas)
L/E - list enumerate (very rural)

Summary

The missing data levels for the 1988 PES were low and reflect on the quality work done in the field and processing offices. The missing P-sample match rates are likely to be higher for the 1990 PES then the 1988 PES since many movers treated as out-of-scope for 1988 will be in-scope for 1990. Mover matching is a difficult operation which will be compounded for 1990 due to shipments of materials between processing offices. For 1990, we plan on producing a measure of error due to imputation which will be added to the sampling error. The estimated imputation error follows the work by Schenker (1989). If similar low rates of missing data can be observed for the 1990 PES, then the errors from missing data can be minimized.

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

References

Childers, Danny and Howard Hogan (1990). "The 1988 Dress Rehearsal Post-Enumeration Survey: Final Results", paper to be presented at the annual meetings of the American Statistical meeting, Anaheim.

Schenker, Nathaniel (1988). "Handling Missing Data in Coverage Estimation, with Application to the 1986 Test of Adjustment Related Operations". Survey Methodology, 14, 87-97

Schenker, Nathaniel (1989). "The Use of Imputed Probabilities for Missing Binary Data". Proceeding of the Fifth Annual Research Conference, US Bureau of the Census, p.133-139