# MISSING DATA IN THE U.S. CENSUS 2000 DRESS REHEARSAL - AN OVERVIEW

Steven P. Hefter, Lisa D. Fairchild, Philip M. Gbur, U.S. Census Bureau
Steven P. Hefter, U.S. Census Bureau, Washington, DC 20233

**Key Words: Data Quality, Item Nonresponse, Imputation, Allocation**

## I. Introduction

The U.S. Census Bureau conducted the Census 2000 Dress Rehearsal (DR) in 1998 in Sacramento, CA; Menominee, WI; and Columbia, SC and surrounding counties. In the Columbia site we used components of a traditional census methodology which included a post-enumeration survey (PES). The DR PES was similar in design to the Integrated Coverage Measurement (ICM) Survey used in the Sacramento and Menominee sites where a sampling census methodology was employed. As with any census or survey, missing data was encountered throughout the process. This paper gives a brief overview of census operations including the initial phase, the ICM/PES, and the estimation methodology and the levels of missing data encountered.

## II. Initial Phase

### A. Operations

The Initial Phase operations included creating a list of the addresses in the three sites; an enumeration of households, group quarters and persons without a usual residence; followup for nonresponse; and the data capture. In each site, people in all residential addresses were given an opportunity to mail back a questionnaire. Those who did not mail back a questionnaire by the cutoff date were included in nonresponse followup (NRFU). Traditionally, enumerators are sent out to all nonresponding households, but sampling for NRFU was done in Sacramento for blocks not selected in the ICM sample. All nonresponding housing units (HUs) in ICM blocks were followed up in the field [1]. This paper discusses data collected by the nonresponse followup enumerators but does not include estimation of nonsampled nonrespondents [2].

### B. Missing Data Procedures

Even after completion of NRFU, sample or otherwise, census questionnaires may still have missing data.

---

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

A questionnaire may not have enough information to determine whether or not the HU is occupied. These are defined to be unclassified returns. Unclassifieds were included in NRFU estimation and thus are not discussed here. The methodology in [2] includes estimation for unclassifieds.

Classified returns must have at least a HU status (occupied, vacant, or delete from the universe) and, if the HU is occupied, the number of persons in the household. Other questionnaires may be complete except for a few questions, referred to as item nonresponse. In addition, as part of the data processing, the questionnaire responses are subject to edits which may result in a response being changed. The assignment of values to complete the questionnaire is called allocation. The allocation procedures used include the following: 1) determination of an appropriate response based on other reported data (such as calculating age from date of birth); 2) assignment of a value based on characteristics of people with similar values for related characteristics (hot deck imputation); and 3) complete substitution of data for a person or all people in a household from a nearby person or household.

If the amount of data provided for an individual is sufficient, then the person is coded as data defined. For the Dress Rehearsal, a person in a HU had to have at least two of the following characteristics to be considered data defined: name, relationship, sex, date of birth or age, Hispanic origin, or race. One or more persons in a household may not be data defined. If all persons in a HU are not data defined then the whole household is substituted [3].

### C. Results

#### 1. Allocation

Allocation percentages are provided in Tables 1-3 by DR site for all people (includes persons in HUs, persons in group quarters, and persons without a usual residence).

Allocation percentages for sex are relatively low and assignment of values was often based upon reported data. The percentages range from 5.0 for South Carolina to 7.1 in Menominee. There is little difference in the percentages across each of the three sites.

The age allocation percentages are quite high, but age could often be determined from reported data

(particularly, date of birth). The percentages range from 25.4 for Sacramento to 41.4 for Menominee but, excluding cases where age was calculated from date of birth, the percentages range from 6.7 for South Carolina to 8.7 for Sacramento. Since age is used only in groupings for estimation, any methodology that is reasonable should have minimal impact on estimation. There is little difference in the latter percentages across the three sites.

The Hispanic Origin allocation percentages range from 8.9 for South Carolina to 10.4 percent for Menominee. However, this allocation is not just for Hispanic/NonHispanic but also includes type of Hispanic. These percentages are also similar for each site.

Race allocation percentages for all people range from 4.6 for South Carolina to 10.2 for Sacramento. The race allocation percentages are particularly high for the Sacramento Hispanic poststratum. They run from 25 to 30 percent. However, since the poststratum assignment is dictated by the Hispanic origin item, this would not affect estimation.

**2. Data Defined Percentages**

Percentages of non-data defined persons are provided in Table 4 by DR site for all people. The total percentages range from 3.4 for South Carolina to 5.4 in Menominee. One or more people in a household may be non-data defined (within household) or all people in a household may be non-data defined (whole household). Of these two types, the within household percentage is higher in Sacramento and Menominee while the percentages of within household and whole household are the same in South Carolina.

Data defined percentages are provided in Tables 5-7 by poststratum race and DR site for all people. The "Non-Data Defined: Imputed" columns include non-data defined persons within households. The "Non-Data Defined: Substituted" columns include persons in whole household substitutions. Comparisons of the data defined percentages with the substituted percentages provide a measure of the results for the whole household substitutions while comparisons of the data defined and total percentages provide an indication of the overall effect of the whole person and whole household imputations or substitutions.

Overall, the largest percentage point differences were seen in South Carolina where 57.3 percent of data defined people are White/Other while only 46.1 percent of persons in whole household substitutions were White/Other with corresponding balancing changes in the percent Black. However, since only 1.7 percent of all people were allocated via whole household substitutions, the differences between the total and data defined person percentages of White/Other and Black are 0.6 percentage points.

Shifts in the distributions between data defined and whole household substituted people are also seen in Sacramento and Menominee. In Sacramento, there is a decrease in the percentage White/Other and Asian with increases in percent Black and Hispanic for the substituted people. For Menominee, there is a decrease in the percentage American Indian with an increase in percent White/Other and Hispanic.

The impact of the somewhat skewed racial distributions of the substituted persons has minor impact on the site-level racial distributions. However, the impact could be pronounced at smaller geographic levels.

**D. Conclusions**

In general, the missing data procedures for the initial phase should not have a significant effect on the site-level estimation. Further research is needed to investigate the effect at smaller geographic levels.

## III. ICM/PES

**A. Sample Design**

The Dual System Estimation (DSE) methodology used in the DR to measure and account for coverage error requires a second, independent enumeration of persons. All addresses in blocks selected for inclusion in the ICM/PES were included in an address listing separate from the census address listing. A cluster sample of HUs was taken from groups of blocks formed for the ICM/PES [4]. Large and small block clusters were subsampled at varying rates to balance field work concerns with sample size considerations. A total of 1,085 block clusters containing 34,890 HUs were included in sample for the DR ICM/PES. This sample is called the P-Sample. The results from the initial phase in these same block clusters are used in conjunction with the P-Sample, and is referred to as the enumeration or E-Sample.

**B. Data Collection and Processing**

Enumerators were sent to the clusters in sample. The HUs listed in these clusters during the independent listing made up the P-Sample. The initial phase HUs in the same clusters made up the E-Sample. The P and E-Sample persons went through a matching process to determine whether the P-Sample persons were residents

on census day and whether they match to the E-Sample. For E-Sample persons it was determined whether they were correctly or erroneously enumerated in the initial phase of the census [5].

### C. Missing Data Procedures

#### 1. Overview

The ICM/PES missing data system was independent from initial phase missing data procedures and accounted for noninterviews and imputed missing responses. The ICM/PES missing data system also calculated a residence probability (for persons with an unresolved residence status) and a match probability (for those persons with an unresolved match status) [6].

#### 2. Whole Household Noninterviews

The ICM/PES missing data system accounted for whole household noninterviews (NIs) in the P-Sample with a NI adjustment. This NI adjustment proportionally redistributed the P-Sample weights of the noninterviewed HUs to the interviewed HUs within block cluster and Type of Basic Street Address.

#### 3. P-Sample Person Characteristic Imputation

The DSE methodology used in the DR required each person in the P-Sample to have sufficient data to be placed in a poststratum. The variables eligible for imputation for P-sample people were race, Hispanic origin, sex, tenure, and age. In the following discussion, the term "previous" refers to a household processed prior to the one in question. Missing tenure was imputed from the closest previous household having the same TOBA. Missing race was imputed from the race distribution within the household. If everyone in the household had a missing value of race, then the nearest previous household, having similar Hispanic origin was used. Hispanic origin was imputed analogously to race. Missing age was imputed from the distribution of age for persons with a similar relationship to, and age of, the reference person. Missing age for single person households was imputed from the age distribution within all single person households. Missing sex was imputed to be the opposite of the spouse's (with spouse present). In households where there was a reference person and a spouse, and both had missing sex, the reference person's sex was imputed from the sex distribution for persons in households where their spouse was present. The spouse's missing sex was then imputed to be the opposite of the reference person's. For all other persons with missing sex, the sex distribution within similar households was used to impute sex for the reference person [6].

#### 4. E-Sample Enumeration Status

The DSE methodology requires that all E-Sample persons be classified as either correctly or erroneously enumerated. E-Sample people with an unresolved enumeration status were assigned a probability of being correctly enumerated in the census. This probability was computed within DR site and before-follow-up match code group as the simple weighted proportion of correct enumerations (among those persons with a resolved final enumeration status).

### D. Estimation Methodology

The person matching results are used in calculating the DSEs for each of the 84 poststrata (6 Race × 7 Age/Sex × 2 Tenure) in each DR site. The DSEs are then placed into a two dimensional matrix (Race by Age/Sex × Tenure). These cell counts are summed to form the marginal constraints. The initial phase estimates for each poststrata were placed in the interior cells of the matrix and the iterative proportional fitting methodology, commonly referred to as raking, was used to force the initial phase estimates to the marginal totals, minimizing the variances and providing results that were consistent across poststrata [7]. The raked DSEs were divided by the initial phase results to yield 84 coverage factors (one for each poststratum) which were used in the subsequent small area estimation [8].

The ICM results were used in the Sacramento and Menominee sites not merely as an evaluation tool, but were incorporated into the final DR estimates. In South Carolina, the official counts did not include the PES results.

### E. Results

#### 1. Noninterview Percentages

A measure of quality for any survey is the number of NIs. However, depending upon the purpose or use of the number, there may be multiple definitions for the percent NI, and this is the case with the DR ICM/PES. The percentages for the DR could vary by three criteria: 1) treatment of vacant HUs; 2) treatment of HUs with a preliminary outcome code of "10"; and 3) use of preliminary versus final outcome codes. For purposes of field operations, vacant HUs are included in the denominator of the percent NI. However, for population estimation purposes, vacants are excluded from the denominator. A preliminary outcome code of "10" represents "No census day residents." In Menominee, this code was erroneously applied to many seasonal vacant HUs that should have received a code of "11" ("Vacant on census day"). Thus, it was decided to convert HUs with an outcome code of "10" to "11." To

ensure consistency across sites, and since the same error may have occurred in the other sites, the conversion was applied to all three DR sites. Note that during final outcome code processing, some additional "10"s may be created - these are left as code "10".

Table 8 includes the components used in calculating the noninterview percentages by site and reflects the final treatment of HUs with a preliminary outcome code of "10". In Menominee, the misclassification of seasonally vacant HUs as noninterviews, if left unaddressed, would have substantially misrepresented the quality of the ICM survey in the site. The final estimation NI percent for Menominee would be 30.9 when including the 176 units with preliminary outcome codes of "10". After these units were reclassified as vacant, the final NI percentage drops significantly to a rate of 1.7. The final estimation NI percentages for Sacramento and South Carolina were 5.1 and 5.2 respectively.

**2. P-Sample Characteristic Imputation Percentages**

Missing data rates for P-Sample persons by DR site are given in Table 9. The selected variables were eligible for allocation. The item nonresponse percentages range from a high of 2.2 percent for age and race in Sacramento to a low of 0.0 percent for Hispanic origin in Menominee. As is to be expected with a coverage survey such as the ICM, all levels of selected item nonresponse were generally very low. Over all three sites, sex had the lowest allocation percentages ranging from 0.1 in Menominee to 0.4 in Sacramento and South Carolina. The high response rate can most likely be attributed to the ease with which the enumerator can determine sex during the interview.

Race and age item nonresponse percentages are relatively consistent across all three sites. Roughly 2.0 percent of all P-Sample persons were allocated these variables. Tenure allocation was also fairly uniform and exceedingly rare in the DR and ranged from 0.2 percent to 0.6 percent. Hispanic origin item nonresponse was generally very low. In Menominee 0.0 percent of the P-Sample people were allocated Hispanic origin. Roughly 1.0 percent of the P-Sample people in Sacramento and South Carolina were allocated Hispanic origin data item.

**3. E-Sample Enumeration Status Percentages**

Table 10 provides the final E-Sample unweighted percentages for each of the four enumeration status match code groups by site. The four groups are correct enumerations, unresolved erroneous enumerations, erroneous enumerations, and persons with insufficient information for matching. The correct enumeration percentages across all three sites were fairly consistent ranging from 86.4 in South Carolina to 88.5 percent in Menominee. The Sacramento site had the highest percentage of people with unresolved enumeration status (3.6) and insufficient information (3.7). The percentages of people with unresolved enumeration status and insufficient information ranged from 0.5 in Menominee to 3.6 in Sacramento, and 1.2 in Menominee to 3.7 in Sacramento, respectively. Menominee had a high rate of erroneous enumerations at 9.7 percent, possibly due to a high incidence of geocoding error. The range of percentages for persons with an erroneous enumeration match code is 6.2 to 9.7.

**F. Conclusions**

As with the initial phase, the handling of missing data did not have a large impact on the ensuing site-level estimation. The amount of missing P-Sample data was small relative to the number of HUs and people interviewed.

In tables 1-10 percentages may not sum due to rounding and the following abbreviations may appear:

NA - Not Applicable
AI - American Indian/Alaska Native
NH/PI - Native Hawaiian/Pacific Islander
Hisp - Hispanic origin
HH - Household

**Table 1:** Sacramento Allocation Percentages by Characteristic and Allocation Component

| Data Item | Base | Percent | Allocation Percent Components | | |
|---|---|---|---|---|---|
| | | | Reported Data | Hot Deck / Consistency | Substituted |
| Sex | 354105 | 6.8 | 1.4 | 3.2 | 2.2 |
| Age | 354105 | 25.4 | 16.8 | 6.5 | 2.2 |
| Hispanic Origin | 354105 | 9.8 | 0.6 | 7.0 | 2.2 |
| Race | 354105 | 10.2 | NA | 8.0 | 2.2 |

**Table 2:** South Carolina Allocation Percentages by Characteristic and Allocation Component

| Data Item | Base | Percent | Allocation Percent Components | | |
|---|---|---|---|---|---|
| | | | Reported Data | Hot Deck / Consistency | Substituted |
| Sex | 647896 | 5.0 | 1.4 | 1.9 | 1.7 |
| Age | 647896 | 28.5 | 21.8 | 5.0 | 1.7 |
| Hispanic Origin | 647896 | 8.9 | 0.2 | 6.9 | 1.7 |
| Race | 647896 | 4.6 | NA | 2.9 | 1.7 |

**Table 3:** Menominee Allocation Percentages by Characteristic and Allocation Component

| Data Item | Base | Percent | Allocation Percent Components | | |
|---|---|---|---|---|---|
| | | | Reported Data | Hot Deck / Consistency | Substituted |
| Sex | 4535 | 7.1 | 1.3 | 3.9 | 1.9 |
| Age | 4535 | 41.4 | 34.0 | 5.6 | 1.9 |
| Hispanic Origin | 4535 | 10.4 | 0.1 | 8.4 | 1.9 |
| Race | 4535 | 6.0 | NA | 4.1 | 1.9 |

**Table 4**: Non-Data Defined Person Percents by Dress Rehearsal Site and Type

| Dress Rehearsal Site | Base | Non-Data Defined Person Percent | | |
|---|---|---|---|---|
| | | Total | Within HH | Whole HH |
| Sacramento | 354105 | 5.1 | 2.9 | 2.2 |
| South Carolina | 647896 | 3.4 | 1.7 | 1.7 |
| Menominee | 4535 | 5.4 | 3.6 | 1.9 |

**Table 5:** Sacramento Data Defined Percents by Poststrata Race

| Race | Total | | Data Defined | | Non-Data Defined: Imputed | | Non-Data Defined: Substituted | |
|---|---|---|---|---|---|---|---|---|
| | # | % | # | % | # | % | # | % |
| Total | 354105 | 100.0 | 336088 | 100.0 | 10392 | 100.0 | 7625 | 100.0 |
| White/ Other | 157630 | 44.5 | 152279 | 45.3 | 2531 | 24.4 | 2820 | 37.0 |
| Black | 54953 | 15.5 | 51547 | 15.3 | 1780 | 17.1 | 1626 | 21.3 |
| AI | 10943 | 3.1 | 10343 | 3.1 | 286 | 2.8 | 314 | 4.1 |
| NH/PI | 2482 | 0.7 | 2294 | 0.7 | 143 | 1.4 | 45 | 0.6 |
| Asian | 57115 | 16.1 | 53116 | 15.8 | 3097 | 29.8 | 902 | 11.8 |
| Hisp | 70982 | 20.0 | 66509 | 19.8 | 2555 | 24.6 | 1918 | 25.1 |

**Table 6:** South Carolina Data Defined Percents by Poststrata Race

| Race | Total | | Data Defined | | Non-Data Defined: Imputed | | Non-Data Defined: Substituted | |
|---|---|---|---|---|---|---|---|---|
| | # | % | # | % | # | % | # | % |
| Total | 647896 | 100.0 | 625804 | 100.0 | 11246 | 100.0 | 10846 | 100.0 |
| White/ Other | 367612 | 56.7 | 358808 | 57.3 | 3805 | 33.8 | 4999 | 46.1 |
| Black | 257772 | 39.8 | 245389 | 39.2 | 6928 | 61.6 | 5455 | 50.3 |
| AI | 3520 | 0.5 | 3366 | 0.5 | 97 | 0.9 | 57 | 0.5 |
| NH/PI | 382 | 0.1 | 369 | 0.1 | 7 | 0.1 | 6 | 0.1 |
| Asian | 6209 | 1.0 | 5948 | 1.0 | 124 | 1.1 | 137 | 1.3 |
| Hisp | 12401 | 1.9 | 11924 | 1.9 | 285 | 2.5 | 192 | 1.8 |

**Table 7:** Menominee Data Defined Percents by Poststrata Race

| Race | Total | | Data Defined | | Non-Data Defined: Imputed | | Non-Data Defined: Substituted | |
|---|---|---|---|---|---|---|---|---|
| | # | % | # | % | # | % | # | % |
| Total | 4535 | 100.0 | 4288 | 100.0 | 163 | 100.0 | 84 | 100.0 |
| White/ Other | 581 | 12.8 | 561 | 13.1 | 6 | 3.7 | 14 | 16.7 |
| Black | 4 | 0.1 | 4 | 0.1 | 0 | 0.0 | 0 | 0.0 |
| AI | 3831 | 84.5 | 3617 | 84.4 | 152 | 93.3 | 62 | 73.8 |
| NH/PI | 1 | 0.0 | 1 | 0.0 | 0 | 0.0 | 0 | 0.0 |
| Asian | 2 | 0.0 | 2 | 0.0 | 0 | 0.0 | 0 | 0.0 |
| Hisp | 116 | 2.6 | 103 | 2.4 | 5 | 3.1 | 8 | 9.5 |

**Table 8**: ICM/PES Noninterview Components based on Final Outcome Codes by Site

| Component | Sacramento | Menominee | South Carolina |
|---|---|---|---|
| Total Addresses | 16419 | 794 | 17677 |
| A. Interview | 14322 | 409 | 14972 |
| B. NI - refusal, no one home, etc. | 486 | 2 | 495 |
| C. NI - no data defined people | 93 | 0 | 66 |
| D. No census day residents (10) | 186 | 5 | 261 |
| E. Vacant | 1118 | 368 | 1208 |
| F. Not a HU | 214 | 10 | 675 |
| Final Estimation NI Rate* | 5.1 | 1.7 | 5.2 |

* NI Rate Definition: (B + C + D) / (A + B + C + D )

**Table 9:** ICM/PES P-Sample Person Missing Data Percentages for Selected Variables by Site

| Variable | Sacramento | Menominee | South Carolina |
|---|---|---|---|
| Base * | 36336 | 1271 | 35920 |
| Sex | 0.4 | 0.1 | 0.4 |
| Age | 2.2 | 1.6 | 2.2 |
| Hispanic Origin | 1.3 | 0.0 | 1.1 |
| Race | 2.2 | 1.0 | 1.8 |
| Tenure | 0.6 | 0.2 | 0.5 |

* Base excludes people with a residence status code of "Remove."

**Table 10:** ICM/PES E-Sample Final Unweighted Correct Enumeration, Unresolved Enumeration Status, Erroneous Enumeration, and Insufficient Information Counts and Percents by Site

| Status | Sacramento | Menominee | South Carolina |
|---|---|---|---|
| Base | 35806 | 1202 | 33959 |
| Correct Enumeration Percentage | 86.5 | 88.5 | 86.4 |
| Unresolved Erroneous Enumeration Status Percentage | 3.6 | 0.5 | 2.4 |
| Erroneous Enumeration Percentage | 6.2 | 9.7 | 9.1 |
| Insufficient Information Percentage | 3.7 | 1.2 | 2.1 |

REFERENCES


[1] Singh, R., Cantwell, P. and Kostanich, D. "Census 2000 Dress Rehearsal Methodology and Results," Presented at the American Statistical Association Joint Statistical Meetings, August 10, 1999.

[2] U.S. Census Bureau, "Specifications for Nonresponse Followup Sampling and Undeliverable-as-Addressed Vacant Sampling for the Census 2000 Dress Rehearsal," internal memorandum for Lynch from Singh, DSSD Census 2000 Dress Rehearsal Memorandum Series #A-35, April 10, 1998.

[3] U.S. Census Bureau, "1998 Dress Rehearsal Census 2000 One Hundred Percent Imputation Specifications, Version 3," Population Division internal document, September 3, 1998.

[4] U.S. Census Bureau, "Computer Specifications for the Selection of the ICM Sample for the Census 2000 Dress Rehearsal", internal memorandum for Lynch from Kostanich, DSSD Census 2000 Dress Rehearsal Memorandum Series #A-5, November 14, 1997.

[5] U.S. Census Bureau, "The Design of the Census 2000 Dress Rehearsal Integrated Coverage Measurement", internal memorandum for Ramos from Childers, DSSD Census 2000 Dress Rehearsal Memorandum Series, Chapter F-DT-2, December 15, 1998.

[6] Ikeda, M and Kearney, A., "Handling of Missing Data in the Census 2000 Dress Rehearsal Integrated Coverage Measurement Sample", Presented at the American Statistical Association Joint Statistical Meetings, August 10, 1999.

[7] Schindler, E., "Iterative proportional fitting in the 2000 Census Dress Rehearsal", Presented at the American Statistical Association Joint Statistical Meetings, August 10, 1999.

[8] U.S. Census Bureau, "Census 2000 Dress Rehearsal Computer Specifications for the Integrated Coverage Measurement Block Level Estimation," internal memorandum for Stoudt from Griffin, DSSD Census 2000 Dress Rehearsal Memorandum Series #A-86.