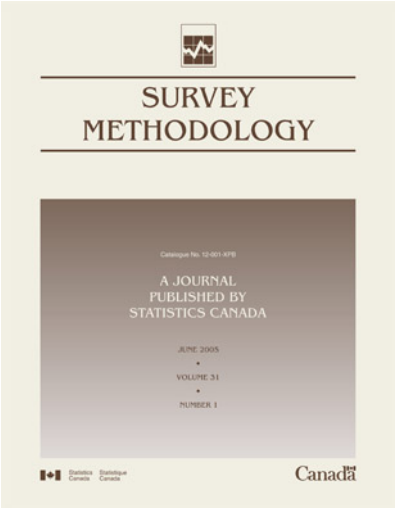
Catalogue no. 12-001-XIE

Survey Methodology

December 2003



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

December 2003

Published by authority of the Minister responsible for Statistics Canada

February 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 29, No. 2, pp. 139-153
Statistics Canada, Catalogue No. 12-001

Handling Missing Data in the 2000 Accuracy and Coverage Evaluation Survey

Patrick J. Cantwell and Michael Ikeda [1]

Abstract

The Accuracy and Coverage Evaluation survey was conducted to estimate the coverage in the 2000 U.S. Census. After field procedures were completed, several types of missing data had to be addressed to apply dual-system estimation. Some housing units were not interviewed. Two noninterview adjustments were devised from the same set of interviews, one for each of two points in time. In addition, the resident, match, or enumeration status of some respondents was not determined. Methods applied in the past were replaced to accommodate a tighter schedule to compute and verify the estimates. This paper presents the extent of missing data in the survey, describes the procedures applied, comparing them to past and current alternatives, and provides analytical summaries of the procedures, including comparisons of dual-system estimates of population under alternatives. Because the resulting levels of missing data were low, it appears that alternative procedures would not have affected the results substantially. However some changes in the estimates are noted.

Key Words: Cell Imputation; Noninterview Adjustment; Logistic Regression; Dual-System Estimation.

1. Introduction

Following the 2000 Census in the United States, the Census Bureau conducted the Accuracy and Coverage Evaluation (A.C.E.) survey. The survey had two goals: (1) to measure the level of net undercoverage across the nation and in various demographic and geographic domains through a statistical technique called dual-system estimation, and (2) to produce revised population counts that could be used to adjust for this net undercoverage – if the adjusted numbers were deemed to be more accurate than the initial census counts (Hogan 2003).

In the process of interviewing and following up respondents in the A.C.E., some households were missed, and certain information needed to calculate the dual-system estimates was not collected from other sample respondents. This paper describes the levels of missing data, discusses the procedures used to address the problem in the A.C.E., and provides some results and evaluations. It should be noted that the term "missing data" applies after all follow-up attempts were made in the field. These activities included multiple attempts at interviews, the use of highly trained clerks and technicians to resolve cases, and the follow-up of cases where a second interview could provide additional required information.

The A.C.E. realized three main types of missing data. First, some households were not interviewed because they could not be contacted or the interview was refused. What makes the situation different in the A.C.E. is that to each sample housing unit, *two* noninterview adjustments were applied; one corrected for noninterviews on Census Day, while the other corrected for noninterviews on the day of the A.C.E. interview. As will be shown, the need for two adjustments reflects the different ways that out-movers and in-movers were treated in the dual-system estimation.

The second type of missing data occurred when information for a household or person was available but specific demographic characteristics needed for dual-system estimation were not collected. For missing tenure (owner vs. non-owner), race, and Hispanic origin, a form of nearest-neighbor hot-deck imputation was used to take advantage of the correlations often found among people living in geographic proximity. In general, the values of age and sex are geographically less clustered, but often well predicted by specific conditions, such as the person's relationship (*e.g.*, spouse, child) to the household's reference person, or whether information is available on the person's spouse. Therefore, national donor distributions conditioned on relevant covariates were used to impute for age and sex. Because characteristic imputation for the A.C.E. was similar to that done in the Post-Enumeration Survey following the 1990 Census, the methods and results are not discussed further in this paper.

The third type also involved item missing data. For a small number of people in the A.C.E., not enough information was collected to determine the resident status (whether or not the person was living in the sampled block cluster on Census Day) or the match status (whether or not the person actually matched to someone in the census). Similarly, some people counted in the census lacked sufficient information to determine whether they were correctly enumerated. The status in such cases is said to be "unresolved." Yet this information is required to compute dual-system estimates. To resolve such cases, a probability of resident (or match or correct enumeration) was assigned as the average weighted value from a set of resolved cases with similar characteristics.

Some of these procedures – described in greater detail below – were applied in similar forms in the 1990 Post-Enumeration Survey, as well as in tests conducted during

1. Patrick J. Cantwell and Michael Ikeda, Mathematical Statisticians, U.S. Census Bureau, Statistical Research Division, Washington, D.C. 20233-9100.

the 1990s. The main exception is the assignment of a probability in the case of unresolved resident, match, or enumeration status. In the Post-Enumeration Survey and at times for specific tests in the 1990s, these probabilities were computed using a logistic regression model. The method applied in the 2000 A.C.E. used less information than some alternatives such as logistic regression, but was simpler to implement and verify in the tight A.C.E. schedule.

The levels of missing data in the A.C.E. were relatively low, which helped to reduce the potential for additional error in the estimates.

- The household noninterview rates were 3.0% and 1.1% (unweighted), respectively, on Census Day and Interview Day.
- The imputation rates for the five A.C.E. characteristics required for dual-system estimation ranged from 1.4% to 2.5% (unweighted and weighted).
- Among people in the A.C.E., the rates of unresolved resident and match status were 2.3% and 1.2% (unweighted), respectively; among census enumerations, only 3.0% (2.6% weighted) of the sample had unresolved enumeration status.

When assigning probabilities for unresolved status, the success of the variables used to define imputation cells was mixed. Variables that used information related to an individual's processing in the survey operations discriminated well among cells. However, variables describing the person's demographic characteristics appear to have been generally less successful.

Section 2 contains background information about the A.C.E. and dual-system estimation. The A.C.E. noninterview adjustment is discussed in section 3. For persons with unresolved resident, match, or enumeration status, a probability was assigned according to procedures described in section 4. Section 5 examines the effect of some alternatives to the A.C.E. missing data procedures on the dual-system estimates of the population. Finally, a few observations are recounted in section 6. For a detailed description of the missing data procedures for the 2000 A.C.E., see Cantwell (2001). Summaries of missing data can be found in Cantwell *et al.* (2001).

In what follows, unweighted frequencies and proportions are generally given. Unless noted otherwise, the weighted numbers are very close. However, the probabilities assigned to unresolved cases in Tables 4, 5, and 6 are the actual weighted ones used in the estimation.

2. A Brief account of the Survey and Dual-System Estimation

Through the Accuracy and Coverage Evaluation (A.C.E.), the Census Bureau attempted to measure and adjust for the historical differential net undercount observed in the U.S. Census (Anderson and Fienberg 1999, page 29). Like the 2000 Census, the A.C.E. covered the entire nation. (A separate sample and analysis were conducted for Puerto Rico.) A sample of about 300,000 housing units in 11,303 block clusters was selected (Fenstermaker 2000, Childers 2000).

To estimate coverage of the population, the A.C.E. relied on dual-system estimation, a method based on capture-recapture methodology (Peterson 1896, Sekar and Deming 1949). Suppose one considers only those housing units contained in the sample of block clusters selected for the A.C.E. After the census enumeration – but without using *any* information collected in the census – the Census Bureau independently interviewed people in the A.C.E sample and obtained a roster of people living in the units on Census Day, April 1, 2000. These results were then matched to (compared with) the census enumeration in those block clusters to estimate how many people were missed. Within the sample block clusters, the units enumerated independently in the A.C.E. were defined as the $P-$Sample, and those enumerated in the census as the $E-$Sample.

In the same sample of block clusters, comparisons and analyses were made to estimate the proportion of census enumerations that were correct, that is, complete, unique, and recorded in the proper location. Erroneous enumerations include people who are duplicated or fictitious, or should not be counted at that address, for example, because their usual residence was elsewhere, such as in a college dormitory. The resulting dual-system estimator is

$$\hat{N} = (C - I)\ \hat{p}_{ce}\left(\frac{1}{\hat{p}_{\text{match}}}\right), \qquad (1)$$

where C is the official census count, including imputed persons and erroneous enumerations; I is the number of whole-person imputations; $\hat{p}_{ce}$ is the weighted estimate of the proportion of correct enumerations in the census; and $\hat{p}_{\text{match}}$ is the weighted estimate of the proportion of $P-$Sample people who match to someone enumerated in the census. People are imputed, for example, when a census enumerator confirms that a certain number of people live at an eligible address, but sufficient additional information cannot be gathered. The actual number of whole-person imputations is known and can be removed from C in the estimate.

Dual-system estimates were calculated separately within population subgroups called post-strata. Post-stratum estimates were then used to determine adjustment factors to be applied to all people counted in the census according to their specific post-stratum. Finally, adjusted counts for any geographic area were calculated by summing the adjusted counts across post-strata in the area. For more detailed information on A.C.E. field operations and dual-system estimation in general, see Childers (2000) and Hogan (1993, 2003), respectively.

3. Noninterview Adjustment

Noninterview adjustment was performed only on the $P-$Sample; in the census (and, thus, in the $E-$Sample), people in all known housing units were accounted for through a variety of procedures. The small number of housing units whose information was collected by a proxy respondent, often a neighbor or building manager, were treated as valid interviews and are not the subject of the noninterview adjustment. Because people moved in and out of housing units between Census Day and the time of the A.C.E. interview, the Census Bureau had to consider the mover status – out-mover, in-mover, or non-mover – of all people in the $P-$Sample, as well as the interview situation at the two different moments. Out-movers were living in the housing unit in question on Census Day, but had moved out before Interview Day. The situation was reversed for in-movers. Non-movers lived in the unit on both days. At the time of the A.C.E. interview, *in one interview* questions were asked to determine who lived in the household on Interview Day and who lived there on Census Day. Mover status was assigned to each person in the sample, and two rosters were created for each household – the Census Day roster and the Interview Day roster.

The A.C.E. used in-movers to estimate the *number* of $P-$Sample movers, while using out-movers to estimate the *match rate* of the movers. The weighted $P-$Sample total, that is, the denominator of $\hat{p}_{\text{match}}$ in equation (2), is estimated as the weighted total of all non-movers and in-movers. Yet the weighted number of $P-$Sample matches is estimated by adding the number of matches among non-movers to the product of the number of in-movers and the match rate for out-movers:

$$\hat{p}_{\text{match}} = \frac{M_{nm} + N_{im} \times \dfrac{M_{om}}{N_{om}}}{N_{nm} + N_{im}}, \qquad (2)$$

where N (people) and M (matches) are indexed by nm, im, and om, representing non-movers, in-movers, and out-movers, respectively. All in-movers and non-movers were generally assumed to be A.C.E. Interview Day residents. (People living in group quarters, such as college students in dormitories, were not eligible for the $P-$Sample.)

The mover procedure used in the A.C.E. differed from that used in the 1990 Post-Enumeration Survey. In 1990 in-movers were used to estimate the number of movers *and* their match rate. For the latter, the in-movers had to be matched back to their address on Census Day. That procedure was changed for the census tests conducted during the 1990's to accommodate the planned use of sampling for census nonrespondents. When the U.S. Supreme Court ruled against the sampling plan in 1999 (*Department of Commerce v. United States House of Representatives*, 525 U.S. 316, 1999), it was thought that changing the mover procedure again so late before the census would introduce unacceptable risks.

Due to the mover procedure described above, each housing unit had two interview statuses – one based on the housing unit's situation as of Census Day, and the other based on the day of the A.C.E. interview. A unit that was vacant, removed from the list of eligible housing units (because, for example, it was demolished or used only as a business), or in certain special places was not considered an interview or a noninterview. Table 1 provides a fictional but illustrative block cluster. It demonstrates how the status of a housing unit on Census Day and Interview Day would have been determined.

Results of the A.C.E. interviewing operation are shown in Table 2. Of the 261,969 housing units occupied on Census Day, 7,794 (3.0 percent) were noninterviews. The corresponding numbers for Interview Day were 267,155 and 3,052 (1.1 percent).

As different interview statuses were possible for a housing unit on Census Day and Interview Day, different noninterview adjustments were required for each day. Each of the two adjustments generally spread the weights of noninterviewed units over interviewed units in the same noninterview cell: the sample block cluster crossed with the type of basic address, defined as single-family, multi-unit (such as apartments and condominiums), or all others. Other characteristics, known for all housing units, could have been used to define the cells. However, the cells were defined to take advantage of the typical local homogeneity, and of the fact that people living in, for example, apartments share many of the characteristics – household size, propensity to move, *etc.* – that are related to capture probabilities in the census.

The noninterview adjustment based on the Census Day status of housing units was used to adjust the person weights of non-movers and out-movers. Similarly, the Interview Day noninterview adjustment was used to adjust the person weights of in-movers. Within a noninterview cell, the adjustment factor for *Census Day* was computed as the weighted sum of interviews and noninterviews for Census Day divided by the weighted sum of interviews for Census Day. A housing unit's weight was the inverse of the final selection probability of its block cluster into the A.C.E. sample. (These weights were trimmed in a very small number of clusters.)

The noninterview adjustment factor for Interview Day was computed as above, but with its status – interview, noninterview, vacant, or delete – being considered for Interview Day rather than for Census Day. The example in Table 1 demonstrates the calculation of the noninterview adjustment for the fictional block cluster. Because the non-interview rates were so small, the noninterview adjustment factors were close to 1 for most housing units in the sample. For Census Day, the factors were less than 1.10 for more than 92% of the units; for Interview Day, the factors were less than 1.10 for over 98% of the units.

Table 1
An Example of Adjustment for Noninterviews

Consider a block cluster with nine housing units, all having the same type of basic address, for example, all single family homes, as depicted below

Housing Unit	Weight	Actual Situation	Status of (and Information from) A.C.E. Interview	Census Day Status	A.C.E. Interview Day Status
1	100	Resident on 4/1/00 and at time of A.C.E. interview	Interviewed in A.C.E.	Interview	Interview
2	100	Resident on 4/1 and at time of A.C.E. interview	Neighbor (proxy) interviewed in A.C.E.	Interview	Interview
3	100	Resident on 4/1 and at time of A.C.E. interview	No one interviewed in A.C.E.	Noninterview	Noninterview
4	100	Vacant on 4/1, resident at time of A.C.E. interview	Interviewed in A.C.E., knows of 4/1 status	Vacant	Interview
5	100	Vacant on 4/1, resident at time of A.C.E. interview	Interviewed in A.C.E., no knowledge of 4/1 status	Noninterview	Interview
6	100	Vacant on 4/1, resident at time of A.C.E. interview	No one interviewed in A.C.E.	Noninterview	Noninterview
7	100	Resident on 4/1, vacant at time of A.C.E. interview	Information obtained from proxy	Interview	Vacant
8	100	Resident on 4/1, vacant at time of A.C.E. interview	No info on 4/1 status; Census staff determines vacant at time of A.C.E.	Noninterview	Vacant
9	100	Resident on 4/1, different resident at time of A.C.E. interview	Interviewed in A.C.E., knows of 4/1status	Interview	Interview

In this noninterview cell (sample block cluster × type of basic address), people in interviewed housing units would have received the following noninterview adjustments:

(1) to the person weights of non-movers and out-movers, the Census Day noninterview adjustment = 800/400 = 2.0

(2) to the person weights of in-movers, the A.C.E. Interview Day noninterview adjustment = 700/500 = 1.4

Table 2
Status of Household Interviews in the A.C.E. (Unweighted)

	Census Day		A.C.E. Interview Day	
	Number	Percent	Number	Percent
Total Housing Units	300,913	100.0	300,913	100.0
Interviews	254,175	84.5	264,103	87.8
Noninterviews	7,794	2.6	3,052	1.0
Vacant Units	28,472	9.5	29,662	9.9
Deleted Units	10,472	3.5	4,096	1.4
Noninterview rate[1]	3.0%		1.1%	

[1] Noninterview rate = Noninterviews/(Interviews + Noninterviews)

When the unweighted number of noninterviewed units in a given noninterview cell was more than twice the unweighted number of interviewed units, the weights of the noninterviewed units in this cell were spread over the interviewed units in a broader set of noninterview cells. This remedy was needed for only 65 cells for the Census Day adjustment, and 13 cells for the Interview Day adjustment. The prescribed procedure differs from the usual collapsing of sparse cells, but allowed us to address such cells in a simple automated fashion. This capability was important under a very tight schedule when it was impossible to predict which cells would have too few interviews. For evaluation purposes, the housing-unit weights were later re-computed under a collapsing scheme, and compared to the weights as determined in the A.C.E. Again, due to the low rates of noninterview, the weights were the same for most units, and close for the rest. The effect on the resulting dual-system estimates is shown in section 5.2.

4. Assigning Probabilities for Unresolved Cases

After all A.C.E. follow-up activities were completed, there remained a small fraction of the A.C.E. sample without enough information to compute the components of the dual-system estimator given in equation (1). Their status was said to be "unresolved."

4.1 Unresolved Cases and their Frequencies

One component of the dual-system estimator in equation (1) is $\hat{p}_{\text{match}}$, the estimated proportion of the $P-$ Sample who match to someone enumerated in the census. In (2) for $\hat{p}_{\text{match}}$, when estimating the number of people (N_{nm}, N_{om}) or matches (M_{nm}, M_{om}) among non-movers and out-movers, only Census Day residents of the sample block clusters were considered; someone who usually lives in a nursing home, for example, was omitted from the computation.

Thus, for each person in the $P-$Sample, determining resident status and match status was required.

After follow-up operations were completed, all people in the $P-$Sample who were eligible to be matched to the census were classified into three types according to their status as a resident in their sampled block *on Census Day:* residents, nonresidents, and unresolved persons – those for whom there was not enough information to determine the resident status. Further, each confirmed or possible (unresolved) Census Day resident in the $P-$Sample was determined to be a match, a nonmatch, or unresolved match. The match status for confirmed Census Day nonresidents, such as in-movers, was not used in the estimation. The estimator in (1) also requires an estimate of the proportion of correct enumerations in the census, $\hat{p}_{ce}$. After whole-person imputations were removed from the $E-$Sample, each remaining person had one of three types of enumeration status: correct, erroneous, or unresolved.

Table 3 summarizes the frequencies of resident and match status in the $P-$Sample, and enumeration status in the $E-$Sample. The table also shows results for nonmovers and out-movers in the $P-$Sample. One can see that the extent of unresolved cases is quite small: 2.3% for resident status, 1.2% for match status, and 3.0% for enumeration status. (The weighted rates are 2.2%, 1.2%, and 2.6%, respectively.) In the 1990 Post-Enumeration Survey, the rate of unresolved matches was 1.9%, and unresolved enumerations was 2.4%. (Resident status was not defined in a manner comparable to 2000.) Care must be taken, however, as the definitions of the several statuses were slightly different in 1990.

4.2 Assigning Probabilities to Unresolved Cases

In the A.C.E., a form of cell imputation was used to assign probabilities for sample cases with unresolved resident, match, or enumeration status. All people in the sample – resolved and unresolved – were placed into groups called imputation cells based on operational and demographic characteristics. Different variables were used to define cells for each type of status. Within each imputation cell the weighted average of 1's and 0's (representing, *e.g.*, match and non-match, respectively) among the resolved cases was calculated, and that average was imputed for all unresolved persons in the cell. More details are provided below.

In the 1990 Post-Enumeration Survey, hierarchical logistic regression was used to calculate probabilities of match and correct enumeration for cases with missing information. (Due to the procedure used to treat movers in 1990, resident status played a different role then.) The model and some results are discussed in Belin *et al.* (1993).

During the 1990s, the Census Bureau originally planned to produce in 2000 adjusted census estimates for each of the 50 states (and the District of Columbia) using data collected only from that state. This approach affected the strategy for treating unresolved status in two ways. First, within each state, there would be far fewer data – resolved cases – on which to build a logistic regression model. Second, there would be 153 different models to examine and verify, separate models for resident, match, and enumeration status in each state. Because the production schedule for the A.C.E. provided only about three weeks for addressing all

Table 3
Final Status Frequencies for the P and $E-$Samples (Unweighted)

		Final resident status			
$P-$Sample	Total people[1]	Confirmed resident	Confirmed nonresident	Unresolved resident	Resident rate for resolved cases
U.S. Total	653,337	95.8%	1.9%	2.3%	98.1%
Mover status					
Non-mover	627,992	96.6%	1.7%	1.7%	98.3%
Out-mover	25,345	75.2%	7.5%	17.4%	91.0%
		Final match status			
$P-$Sample	Total people[2]	Match	Nonmatch	Unresolved match	Match rate for resolved cases
U.S. Total	640,945	90.3%	8.5%	1.2%	91.4%
Mover status					
Non-mover	617,490	91.1%	8.0%	0.9%	91.9%
Out-mover	23,455	67.8%	21.7%	10.5%	75.8%
		Final enumeration status			
$E-$Sample	Total people	Correct enumeration	Erroneous enumeration	Unresolved enumeration	Correct enumeration rate for resolved cases
U.S. Total	704,602	92.6%	4.4%	3.0%	95.5%

[1] Those in the $P-$Sample eligible to be matched to the census.
[2] Confirmed or possible residents in the $P-$Sample.

aspects of missing data, it was believed that a procedure to handle unresolved status that was simpler to implement and verify would reduce the risk of not completing the dual-system estimates under the imposed deadline. Cell imputation provided the desired simplicity, but its accuracy relative to that under logistic regression modeling had to be evaluated in subsequent testing.

During census tests in 1995 and 1996, certain types of unresolved status were addressed using logistic regression, while cell imputation was used for other types. The latter procedure was used exclusively in the Census Dress Rehearsal in 1998 (Ikeda, Kearney and Petroni 1998), when the Census Bureau was still preparing to produce estimates independently within each state. Data from these tests indicated that the exact method of calculating probabilities for unresolved status (match, resident, or correct enumeration) had only a minor effect on the dual-system estimates. Details of this research can be found in Petroni (1997, 1998a, 1998b, and 1998c).During census tests in 1995 and 1996, certain types of unresolved status were addressed using logistic regression, while cell imputation was used for other types. The latter procedure was used exclusively in the Census Dress Rehearsal in 1998 (Ikeda, Kearney and Petroni 1998), when the Census Bureau was still preparing to produce estimates independently within each state. Data from these tests indicated that the exact method of calculating probabilities for unresolved status (match, resident, or correct enumeration) had only a minor effect on the dual-system estimates. Details of this research can be found in Petroni (1997, 1998a, 1998b, and 1998c).

With the decision by the U.S. Supreme Court in 1999 (*Dept. of Commerce v. U.S. House of Representatives*), the Census Bureau changed the design of the survey and removed the restriction that adjusted estimates be based solely on data from within each state. However, there remained concerns about implementing a logistic regression approach that had not been tested in the Dress Rehearsal. Further, there was no guarantee that available software would adequately run logistic models on data sets the size of the entire A.C.E. sample (between 640,000 and 750,000 people). Based on these concerns and research findings on relative accuracy, a decision was made to use the simpler procedure, cell imputation, to resolve missing status in the A.C.E.

To demonstrate how cell imputation was applied in the A.C.E., one can look at resident status; the method was applied analogously to match and enumeration status. First, all non-movers and out-movers in the P Sample were placed into a number of imputation cells according to operational and demographic characteristics, as defined in Table 4; in-movers were left out, as their Census Day resident probability was 0 by definition. *Among the resolved cases* in cell i, denoted by the set $R(i)$, an indicator variable for resident status was defined as $1_{res,\,j}=1$, if person j was a resident in the household on Census Day, or 0, otherwise. Then within cell i, the weighted proportion of Census Day residents, was computed:

$$P(res)_i = \frac{\sum_{j\in R(i)} w_j 1_{res,\,j}}{\sum_{j\in R(i)} w_j} \qquad (3)$$

where w_j is the weight of person j incorporating all stages of sampling. $P(res)_i$ was then assigned to each unresolved person in cell i, that is, each of the 15,082 people (2.3% of 653,337) with unresolved resident status. (The exception is for match code group 7, as explained below.) Table 4 provides the resident probabilities assigned within the cells. This assignment defines for all cases – resolved and unresolved – an "extended" indicator, allowing values between 0 and 1:

$$1'_{res,\,j} = \begin{cases} 1_{res,\,j}, & \text{if } j\in R(i) \\ P(res)_i, & \text{otherwise} \end{cases} \qquad (4)$$

The estimated numbers of non-movers and out-movers in the $P-$Sample in (2), N_{nm} and N_{om}, respectively, are then computed by attaching the person weight and summing the indicator $1_{res,\,j}$ over the non-movers and out-movers, respectively, in all cells. The number of matches, M_{nm} or

Table 4
Imputation Cells for Resolving Resident Status in the $P-$Sample

$P-$Sample	Owner		Non-Owner	
Match Code Group	Nonhispanic White	Others	NonhispanicWhite	Others
1. Matches needing follow-up	0.982	0.986	0.993	0.991
2. Possible matches	0.973	0.968	0.966	0.972
3a. Partial household nonmatches needing follow-up; aged 18-29, child of reference person	0.755	0.901	0.883	0.928
3b. Partial household nonmatches needing follow-up; others not in 3a	0.956	0.971	0.959	0.969
4. Whole household nonmatches needing follow-up, not conflicting households	0.920	0.943	0.911	0.914
5. Nonmatches from conflicting households	0.910	0.927	0.945	0.954
6. Resolved before follow-up	0.993	0.990	0.990	0.988
7. Insufficient information for matching (Weighted column average of groups 1-5 and 8)	0.813	0.867	0.844	0.872
8. Potentially fictitious or said to be living elsewhere on Census Day	0.119	0.123	0.177	0.157

M_{om}, and thus, $\hat{p}_{\text{match}}$ are determined analogously, as is $\hat{p}_{ce}$, in the case of enumeration status.

In the Census Dress Rehearsal of 1998, cell imputation for unresolved resident probability was used with only three cells: persons sent to follow-up, persons not needing follow-up, and persons with insufficient information for matching. For the third cell, which contained no resolved cases, a proportion based on all resolved cases in the first two cells was assigned. Results from the Dress Rehearsal (Kearney and Ikeda 1999) suggested that dividing the $P-$Sample into the various match code groups would be helpful. Further research and discussion suggested adding other demographic variables within match code group. The larger A.C.E. sample size in 2000 made it possible to support a larger set of imputation cells.

For the A.C.E. in 2000, match code groups 1 through 7 were determined from the match codes and other variables derived *before* the follow-up operation, as explained in Childers (2000). Group 8 was formed differently. Some information from the follow-up operation was coded in time for the A.C.E. missing data procedures. (Under the original schedule, this information would have become available too late to be of use.) *After* the follow-up operation, a small number of people in the $P-$Sample were coded as being potentially fictitious or said to be living elsewhere on Census Day. Among the resolved cases in this group, the probability of being a resident was much lower than for resolved people in other groups. Thus, people satisfying the conditions for group 8 were placed there first, and each of the remaining people was placed appropriately in one of the first seven groups.

Two tenure categories were used: owners and non-owners. Persons were also placed into one of two race-ethnicity categories: Nonhispanic white, and all others. People of multiple races (for example, a person responding as White and Asian) were placed in the latter group. Match code group 3, partial household nonmatches, was split into two subgroups. The first, 3a, included persons in group 3 who were 18 to 29 years of age and were listed on the A.C.E. household roster as a child of the reference person. These were young people many of whom were attending college, sharing residence with colleagues, or moving in and out of their parents' residence. Classification and regression tree analysis, applied to data from the Census Dress Rehearsal of 1998, suggested that this combination of characteristics would discriminate well with respect to resident status. The group 3b included all other persons in group 3.

The resident probability for unresolved $P-$Sample persons was computed as described above, except for those in match code group $7-$ people with insufficient information for matching. Within this row in Table 4, there were essentially no resolved cases from which to extract a probability of being a Census Day resident. Because of their lack of information – most of these cases did not even have a valid name – these people did not go through the matching operation and were not sent to follow-up. To determine a resident probability for these cases, a weighted proportion of Census Day residents (1's and 0's) was computed among the resolved cases in match code groups 1 through 5 and 8, separately for each of the four tenure × race/ethnicity classes. This probability was then assigned to those in group 7. Left out of this computation were those people who were resolved before follow-up (group 6). Observations from the Dress Rehearsal indicated that, in terms of their demographic and operational characteristics, people in group 7 tended to be more like those in groups 1-5 and 8, than like those in group 6.

The issue of unresolved matches was treated like that for unresolved resident status in (3) and (4), with resident status replaced by match status, but with a different set of cells, as is seen in Table 5. Confirmed nonresidents were excluded from the computations of match probabilities.

For unresolved match probability in the Dress Rehearsal, only one imputation cell was used within each of the geographic sites. Subsequent analysis (Kearney and Ikeda 1999) showed that mover status (non-mover *vs.* out-mover) discriminated well between matches and nonmatches among the resolved cases. Thus, for the 2000 A.C.E. mover status was used to define imputation cells for match status. The housing-unit address match code refers to the initial match between housing units on the independent (A.C.E.) listing and the census address list; conflicting housing units, determined during A.C.E. person match activities, were those where the census and A.C.E. rosters had two completely different lists of residents for Census Day (Childers 2000).

It should be noted that 98.3% of the unresolved matches (7,693 of 7,826) were people with insufficient information for matching. As mentioned above, most of them did not have a valid name, and almost all (7,506) were not sent to follow-up. Further, their rate of missing characteristics was much higher than average. Therefore, little useful predictive information was available when forming imputation cells for match status. Variables such as age and ethnicity – that had a higher chance of being imputed and might be of questionable quality – were avoided.

People with at least one imputed demographic variable (age, sex, tenure, race, or Hispanic origin) were grouped when resolving match status. Unpublished studies indicated that – at least among resolved cases in the Dress Rehearsal – the presence of these imputed characteristics is negatively associated with the propensity to be a match. Out-movers from a unit that was a nonmatch or a conflicting household were not separated according to this variable to ensure a reasonable number of resolved cases in each cell from which to estimate the proportion of matches.

In the $E-$Sample, unresolved enumeration status was addressed as discussed above. See Table 6.

Table 5
Imputation Cells for Resolving Match Status in the $P-$ Sample

Mover Status	Housing-Unit Address Match Code: Housing unit was a match		Housing unit was a nonmatch or the household was conflicting	
Non-mover	No imputed characteristics [1] 0.945	1 or more imputed characteristics 0.901	No imputed characteristics 0.690	1 or more imputed characteristics 0.567
Out-mover	No imputed characteristics 0.798	1 or more imputed characteristics 0.791	0.516	

[1] Among the characteristics age, sex, tenure, race, or Hispanic origin.

Table 6
Imputation Cells for Resolving Enumeration Status in the $E-$ Sample

$E-$ Sample Match Code Group	No Imputed Characteristics [1]		1 or More Imputed Characteristics
1. Matches needing follow-up	0.977		0.977
2. Possible matches	0.968		0.968
3a. Partial household nonmatches; aged 18-29, child of reference person	0.871		0.908
3b. Partial household nonmatches; others not in 3a	0.974		0.960
4. Whole household nonmatches where the housing unit matched; not conflicting households	Nonhispanic White	Others	
	0.965	0.974	0.958
5. Nonmatches from conflicting households; for housing units not in regular nonresponse follow-up	0.975		0.965
6. Nonmatches from conflicting households;housing units in regular nonresponse follow-up	0.914		0.926
7. Whole household nonmatches, where the housing did not match in housing-unit matching	Nonhispanic White	Others	
	0.959	0.947	0.950
8. Resolved before follow-up	Nonhispanic White	Others	
	0.995	0.990	0.979
9. Insufficient information for matching	0 (assigned by definition)		
10. Targeted extended search cases[2]	0.928		0.858
11. Potentially fictitious people	0.058		0.088
12. People said to be living elsewhere on Census Day	0.229		0.210

[1] Among the characteristics age, sex, tenure, race, or Hispanic origin.

[2] Targeted extended search refers to a field operation conducted to reduce the variance in the dual-system estimates caused by clustered geocoding errors. For more information, see Navarro and Olson (2001).

As with resident status for $P-$ Sample people, a key factor in determining enumeration status was the $E-$ Sample person's match code group, although the match code groups were defined differently for the two samples. Similar to the $P-$ Sample, people coded as potentially fictitious or said to be living elsewhere on Census Day during the follow-up operation were first placed in groups 11 or 12, respectively. The remainder of the $E-$ Sample was then placed in the appropriate match code group, as defined in the table. Group 3 was split into two subgroups, as when determining resident status in the $P-$ Sample. That is, people aged 18 to 29 who were children of the reference person were separated. Other characteristics used to define cells were the presence or absence of imputed characteristics, as defined in the imputation cells for match status; and whether the person was Nonhispanic white or any other race-ethnicity combination. It should be noted that, according to A.C.E. procedures, anyone in the $E-$ Sample with insufficient information for matching (group 9) was automatically assigned an enumeration probability of 0.

4.3 Comparing Probabilities Under Cell Imputation and Logistic Regression

It can be insightful to compare the probabilities assigned to cases with unresolved status under alternative procedures. Belin (2001) presents such a comparison under a logistic regression model that considered 186 predictors for resident and match status, and 202 predictors for enumeration status. The variables included most of those used in the cell estimation described in section 4.2, as well as individual demographic characteristics, such as age, gender, and relationship to the household's reference person; information about the A.C.E. interview, such as whether the respondent was a proxy; information derived from the sampling process; local-area features, such as whether the area was urban or non-urban; and the interactions among the variables. As the models were fit to the resolved cases sent to follow-up, and then applied to unresolved cases to predict a probability, the models are ignorable in the sense that unresolved status is not considered as a covariate in the underlying model. (See Rubin 1976.)

Tables 7 and 8 summarize the probabilities assigned to unresolved cases under A.C.E. cell imputation and the logistic modeling averaged over the different match code groups. Recall that cell imputation probabilities were computed from weighted data as in (3), while the logistic regression models were run on unweighted data. The predicted probabilities for the two procedures were averaged across all unresolved people unweighted. With an exception to be mentioned later, probabilities and estimates in the A.C.E. were typically similar when using unweighted and weighted data, as the sample was designed to avoid a wide range of weights.

Table 7
Average Resident and Match Probabilities Under Cell Imputation and Logistic Regression

P – Sample Match Code Group	Resident Status			Match Status		
		Avg. Probability Assigned			Avg. Probability Assigned	
	Number Unresolved	Cell Imputation	Logistic Regression	Number Unresolved	Cell Imputation	Logistic Regression
1. Matches needing follow-up	767	0.989	0.983	4	0.848	0.941
2. Possible matches	352	0.970	0.962	131	0.889	0.837
3. Partial household nonmatches	1,306	0.956	0.951	71	0.893	0.050
4. Whole household nonmatches	1,610	0.917	0.926	36	0.770	0.010
5. Nonmatches, conflicting household	1,455	0.940	0.927	49	0.616	0.070
6. Resolved before follow-up	129	0.990	0.990	23	0.842	0.940
7. Insufficient information	7,506	0.844	0.851	7,506	0.835	0.880
8. Fictitious, living elsewhere	2,402	0.148	0.167	6	0.655	0.041

Table 8
Average Enumeration Probabilities Under Cell Imputation and Logistic Regression

E – Sample	Enumeration Status		
		Avg. Probability Assigned	
Match Code Group	Number Unresolved	Cell Imputation	Logistic Regression
1. Matches needing follow-up	711	0.977	0.986
2. Possible matches	305	0.968	0.967
3. Partial household nonmatches	2,191	0.962	0.963
4. Whole household nonmatches where the housing unit matched; not conflicting	4,813	0.967	0.974
5. Nonmatches from conflicting households; housing units not in nonresponse follow-up	532	0.973	0.973
6. Nonmatches from conflicting households; housing units in nonresponse follow-up	779	0.917	0.926
7. Whole household nonmatches, where the housing unit did not match	3,881	0.954	0.961
8. Resolved before follow-up	179	0.990	0.982
9. Insufficient information for matching	0	–	–
10 Targeted extended search cases	2,902	0.918	0.679
11. Potentially fictitious people	1,690	0.064	0.077
12. People said to be living elsewhere on Census Day	3,152	0.225	0.280

Comparing procedures, one sees almost no difference in the average probabilities assigned for resident status. This is not surprising, as cell imputation used the match code group (among other variables) to define cells. Match status presents a different story. To recall, match code group was not used in the cell imputation, as almost all unresolved matches (98.3% of 7,826; 7,506 before the follow-up operation, and 187 more after follow-up) had insufficient information for matching. The first two groups have slightly different probabilities assigned under the two procedures. But in groups 3, 4, and 5, all nonmatches before follow-up, the average probabilities are high under cell imputation (0.893, 0.770, and 0.616), and very low under logistic regression (0.050, 0.010, and 0.070). Of the 156 cases in the three cells, 134 were people each of whom was given an initial code indicating a "nonmatch"; later it was determined correctly that the person had insufficient information for matching. In almost every case, the A.C.E. interviewer recorded a name like "Child Jones", "José Don't Know", or "Unknown Smith". Such cases should have been caught before matching by a clerk, and assigned an initial code of insufficient information. Instead, a match to the census was attempted and failed. If not for this error, such people would have been placed in group 7, where their match probability under logistic regression would have been much higher. Thus, for this small set of 134 cases, the logistic variable, match code group, takes on an incorrect value, and the model predicts a probability – much too low – based on the many resolved cases in group 3, 4, or *5 who really were nonmatches*, but were sent to follow-up primarily to resolve their resident status, not their match status.

The predicted match probabilities in group 8 were also very different. However, with only six unresolved cases, the effect on estimation would be minimal.

Comparing average enumeration probabilities by match code group in Table 8, one sees almost no difference except in group 10, targeted extended search cases. There, the average probability assigned by cell imputation, 0.918, is much higher than that predicted by logistic regression, 0.679. The difference can be explained by the weighting. In the $E-$Sample, of 32,334 people eligible for the targeted extended search operation, 8,298 (all in match code group 10) were sampled out to contain costs and given an A.C.E. weight of 0. The matching operation did not try to determine whether the 8,298 cases were enumerated correctly or not, but simply left them on the data file as erroneous enumerations. Probabilities based on cell imputation were assigned as in equations (3) and (4), incurporating the A.C.E. weight. This removed from the computation those who were sampled out of the A.C.E. The logistic regression model was run on unweighted data and included the 8,298 cases in group 10, bringing down the probability of a correct enumeration predicted for the 2,902 people with unresolved enumeration status.

5. The Effect of Some Alternative Missing Data Procedures on Dual-System Estimates

In the last section, predicted probabilities were compared across two options for treating cases with unresolved status. But the ultimate effect of competing procedures is seen in the resulting dual-system estimates. In this section, several alternatives to those used in the A.C.E. for addressing missing data are compared via the resulting estimates. When they differ significantly, it is not clear which procedure is to be preferred. It should be noted that the A.C.E. estimates released by the U.S. Census Bureau in March of 2001 have been revised following further analyses (Haines 2003). Even though the A.C.E. data are flawed and A.C.E. estimates should generally not be used, it is believed that they are adequate to evaluate the differences in the estimates caused by alternative missing data approaches.

5.1 Results from an Early Evaluation

In the months after initial dual-system estimates from the A.C.E. were released, alternatives to the applied missing data procedures were studied. There were several reasons: estimating the variation that might result from the alternatives, incorporating this variation into total error and loss function analysis for the A.C.E. dual-system estimates, and investigating the viability of non-ignorable missingness procedures for addressing unresolved status. As the results are available in Keathley, Kearney, and Bell (2001), only a summary will be provided here.

Three alternatives involving the noninterview adjustment were examined. The first defined cells differently for the adjustment, adding variables such as race, Hispanic origin, tenure, and household size, as determined from a match to the census file. This procedure tended to produce larger dual-system estimates. Two other noninterview alternatives had no apparent affect on the estimates. In one, a nearest-neighbor noninterview adjustment, the weight of a noninterviewed household was added to that of the nearest interviewed household in the sorted file. In the second, the last 30% of A.C.E. interviews completed were labeled as "late" interviews. The weights of noninterviewed units were added only to the weights of late interviews. These alternatives tried to take advantage of the anticipated homogeneity of units induced by geographic proximity or time of response to the A.C.E.

The other alternatives described in Keathley *et al.* (2001) address unresolved resident, match, or enumeration status. A "late" data approach used information collected only from the last 30% of interviews in the $P-$Sample, or housing units that required nonresponse follow-up in the $E-$Sample. By itself, this approach did not appear to affect the dual-system estimates. The remaining alternatives involved logistic regression models to predict probabilities for

unresolved cases. First, an ignorable logistic model, the one described above (Section 4.3) in Belin (2001), was applied to unresolved resident, match, and enumeration status and tended to produce smaller dual-system estimates (47,481 smaller for the U.S. total). However, it appears that the lowered (on average) enumeration probabilities assigned to the 2902 unresolved cases in the $E-$Sample match code group 10 (see section 4.3) would have more than accounted for this decrease.

Perhaps more interesting are three alternatives that attempted to construct non-ignorable logistic models by lowering the probabilities assigned to unresolved cases, on the premise that ignorable models may overstate the underlying probabilities (Belin 2001). Data from the 1990 Post-Enumeration Survey and its evaluation follow-up were used to estimate non-ignorable effects and incorporate them into the 2000 logistic models. This strategy tended to produce larger dual-system estimates when applied to unresolved match probabilities, and smaller estimates when applied to resident or enumeration probabilities. This result is not surprising, based on equation (1) and the fact that the average match probability assigned to cases with unresolved resident status is less than that for cases with resolved resident status. Although the study's authors conclude that "[t]here is no evidence to suggest that the non-ignorable missingness procedures that we considered are or are not viable alternative missing data procedures" (Keathley *et al.* 2001, page 2), Belin's approach takes a promising step toward addressing the non-ignorability of the missing status.

5.2 Analyses on Other Alternative Procedures

In this section, differences in the dual-system estimates are presented under six numbered alternatives described and motivated below. The results are provided in Table 9 for the U.S. total and for breakdowns by race-ethnicity, tenure, and age. For a precise definition of the race-ethnicity domains, see Kostanich (2001). (Note that a small part of the U.S. population was not part of the A.C.E. universe.) For each alternative, the three numbers given are (a) the difference: the alternative estimate minus the A.C.E. estimate; (b) the standard error of that difference; and (c) the percent relative difference.

Alternative (1) reconsidered the noninterview procedure as applied in the A.C.E. to adjustment cells with a relatively small number of completed interviews. (See section 3.) In this alternative, instead of spreading weights from non-interviewed units over a wider range of cells, cells with too few interviews were collapsed with nearby cells, and noninterview adjustment factors were computed afresh in the newly created cells. Except for Nonhispanic Blacks, none of the estimated differences in Table 9 under this alternative are statistically significant (greater than two standard errors). Similarly, except for several race-ethnicity domains less than two million in size, none of the relative differences are greater than 0.01%.

Alternatives (2), (3), and (4) were derived after examining the effects of the variables used in the imputation cells on the resulting assigned probabilities. From the probabilities assigned in Tables 4 and 6, it is clear that the match code groups discriminated well with regard to resident and enumeration status. Yet it appears that dividing the cells based on demographic variables, such as "Nonhispanic white" *vs*. "Other," made less of a difference. To investigate the effect of demographic variables on the imputation, new probabilities were assigned for unresolved status without using them. Specifically, all resolved and unresolved cases were combined across cells for Nonhispanic white and Other (resident and enumeration status), for match code groups 3a and 3b (resident and enumeration), and for "No imputed characteristics" and "1 or more imputed characteristics" (match and enumeration); the variables derived from A.C.E. operations – match code group, housing-unit address match code, and mover status – were retained. Alternative (2) applies the smaller set of cells only in the $P-$Sample, that is, only for unresolved resident and match status; alternative (3) applies it only in the $E-$Sample (enumeration status); and alternative (4) applies it to both samples.

Under alternative (2), the greatest change in the resident probabilities assigned to unresolved cases occured in the four (original) imputation cells in group 3a, affecting only 96 people with unresolved status. In most other cells for resident status (over 99% of the cases), the probabilities changed very little. A large difference in match probabilities occurred only in the cell "non-mover, nonmatched unit or conflicting household, one or more imputed characteristics," containing 421 unresolved cases. The variable differentiating the number of imputes appears to have had an effect here; if its two "impute" subcells are collapsed, the probability assigned to the "one or more" cell is dominated by the much larger number of resolved people with no imputes, raising the value from 0.567 to 0.684. As is seen in Table 9, the effect on the dual-system estimates is statistically significant for the U.S. total and almost all the breakdowns shown, except for two race-ethnicity groups with sizes under one million people. The relative differences do not appear to be very large, however, ranging from 0.01% to 0.04%. It is not obvious which missing data option produces estimates closer to the unknown true values.

Table 9

Dual-System Estimates Under Alternative Missing Data Procedures

Each cell to the right of the vertical bar contains, in order, estimates of (a) the difference: the alternative estimate minus the A.C.E. estimate, (b) the standard error of that difference, and (c) the relative difference as a percent.

		Estimated Differences Based on Six Alternatives to A.C.E. Missing Data Procedures					
	A.C.E. Estimate (Standard Error)	(1) Noninterview Adjustment With Collapsed Cells	(2) Collapsed Imputation Cells: *P* – Sample Only	(3) Collapsed Imputation Cells: *E* – Sample Only	(4) Collapsed Imputation Cells: *P* and *E* – Samples	(5) Imputing Probabilities Based on the MES	(6) Imputing Probabilities Based on the MER
U.S. Total	276,848,873	−4,299	−55,284	−568	−55,852	−63,632	385,969
	(366,543)	(7,423)	(1,623)	(2,581)	(3,045)	(5,368)	(24,358)
		0.00%	−0.02%	0.00%	−0.02%	−0.02%	0.14%
Race-Ethnicity Domains							
Nonhispanic White	194,226,285	−2,467	−32,324	−1,677	−34,000	−61,817	108,604
	(265,893)	(6,247)	(1,055)	(1,870)	(2,163)	(4,534)	(13,026)
		0.00%	−0.02%	0.00%	−0.02%	−0.03%	0.06%
Nonhispanic Black	34,210,774	−3,495	−11,136	−119	−11,255	−1,303	124,710
	(118,415)	(1,290)	(753)	(1,328)	(1,528)	(1,417)	(11,343)
		−0.01%	−0.03%	0.00%	−0.03%	0.00%	0.36%
Hispanic	35,552,109	725	−8,132	1,432	−6,700	196	124,937
	(138,870)	(3,016)	(857)	(973)	(1,297)	(1,577)	(10,657)
		0.00%	−0.02%	0.00%	−0.02%	0.00%	0.35%
Native Hawaiian or Pacific Islander	618,698	−98	−73	88	15	−107	1,330
	(17,873)	(81)	(72)	(43)	(85)	(74)	(616)
		−0.02%	−0.01%	0.01%	0.00%	−0.02%	0.22%
Nonhispanic Asian	10,056,009	709	−3,175	−257	−3,431	−414	19,556
	(64,372)	(571)	(356)	(439)	(567)	(576)	(3,704)
		0.01%	−0.03%	0.00%	−0.03%	0.00%	0.19%
American Indian on Reservation	567,053	−245	−59	61	2	−38	1,402
	(7,235)	(300)	(49)	(17)	(52)	(73)	(250)
		−0.04%	−0.01%	0.01%	0.00%	−0.01%	0.25%
American Indian *not* on Reservation	1,617,944	572	−386	−96	−482	−148	5,430
	(22,032)	(661)	(68)	(174)	(186)	(144)	(1,446)
		0.04%	−0.02%	−0.01%	−0.03%	−0.01%	0.34%
Tenure							
Owner	188,764,543	−2,237	−34,503	933	−33,570	−7,816	125,058
	(260,408)	(3,805)	(1,205)	(1,971)	(2,317)	(1,942)	(10,063)
		0.00%	−0.02%	0.00%	−0.02%	0.00%	0.07%
Non-Owner	88,084,330	−2,063	−20,782	−1,501	−22,282	−55,816	260,911
	(226,108)	(6,057)	(1,121)	(1,607)	(1,935)	(5,071)	(21,684)
		0.00%	−0.02%	0.00%	−0.03%	−0.06%	0.30%
Age Group							
0–17	73,076,071	2,924	−21,872	−3,315	−25,186	−8,559	107,308
	(137,126)	(2,624)	(625)	(1,324)	(1,474)	(2,047)	(9,785)
		0.00%	−0.03%	0.00%	−0.03%	−0.01%	0.15%
18–49	129,785,393	−2,721	−23,304	3,247	−20,057	−44,534	244,070
	(208,070)	(4,714)	(1,143)	(1,565)	(1,930)	(3,777)	(16,245)
		0.00%	−0.02%	0.00%	−0.02%	−0.03%	0.19%
50 and Over	73,987,409	−4,502	−10,108	−500	−10,608	−10,538	34,591
	(111,125)	(2,766)	(563)	(670)	(877)	(1,421)	(4,561)
		−0.01%	−0.01%	0.00%	−0.01%	−0.01%	0.05%

Under alternative (3), the enumeration probabilities were re-computed using only the match code groups as imputation cells. Noticeable changes were detected in the probabilities in the (original) cells for match code group 3a. In the dual-system estimates, the only significant differences were found in two of the three age categories and some of the small race-ethnicity domains. Except for the latter domains, all the percent differences were under 0.01%. As alternative (4) uses the re-computed probabilities from the P and $E-$Samples, the resulting estimates here were dominated by the $P-$Sample results and thus were similar to those under alternative (2).

The final two alternative procedures employed the same set of imputation cells as those used in the A.C.E., but assigned to unresolved cases in both the P and $E-$ Samples potentially improved probabilities, as determined from one of two evaluations conducted by the Census Bureau following the A.C.E. Alternative (5) secured its probabilities from the Matching Error Study (MES), while alternative (6) based them on the Measurement Error Reinterview (MER). Each study took place in a set of evaluation clusters, a roughly one-in-five subsample of the A.C.E. sample block clusters. Information on the MES and MER sample designs can be found in Petroni (2001) and Killion (2000).

The primary purpose of the MES was to evaluate the A.C.E. person matching operation. The evaluation clusters were rematched by expert matchers, and appropriate changes were made to final match codes and person status. No additional data were collected for the MES. Imputation cell probabilities based on MES data were generally similar to those assigned in the A.C.E. One exception, for resident status, was in the cell for match code group 4, Nonhispanic white, non-owner. Here, the MES probability, 0.712, was much lower than the A.C.E. value of 0.911. This resulted from one cluster in the cell that had 24 persons with large weights geocoded incorrectly, as detected in the MES. The MES enumeration probability for match code group 11, "1 or more imputed characteristics," 0.176, was a bit higher than that for the A.C.E., 0.088. Most other probabilities for resident, match, and enumeration status were close (within 0.03) between the A.C.E. and MES; all others were within 0.07.

In contrast, the MER was designed to evaluate the *data collection* error arising from the A.C.E. matching process. People in the MER were reinterviewed about nine months after Census Day to collect information analogous to that collected in the A.C.E. follow-up operation, but in greater detail. Based on the MER, resident probabilities tended to be substantially higher for the cells in match code group 8, but to be lower for the cells in groups 3, 4, and 5 (denoting nonmatches). The reductions tended to be larger in cells where group 8 took more cases away from groups 3, 4, and 5. One might note that the MER cells in subgroup 3a were fairly small. The "Nonhispanic white, non-owner" cell had only 34 unweighted resolved persons, while the other three cells in group 3a ranged from 125 to 140 unweighted resolved persons. The MER probabilities for enumeration status exhibited similar behavior, with probabilities in groups 11 and 12 raised, and those in the nonmatch groups (3 through 7) lowered. Match probabilities were similar between A.C.E. and MER, mostly differing by 0.01 to 0.05.

Before looking at the dual-system estimates under alternatives (5) (MES probabilities) and (6) (MER probabilities), one should note that, *for the comparison in Table 9*, only the probabilities assigned to unresolved cases were changed based on data collected through the MES or MER. Although the evaluated status of some people may have changed (for example, from nonmatch to match, or confirmed resident to unresolved resident) based on the evaluations, their status was not changed when computing these estimates, as the goal of this exercise was only to explore different methods or information *as they affect the missing data procedures component* of the dual-system estimates.

Under alternative (5), based on MES data and probabilities, the estimates decreased in almost all population domains in Table 9, although never more than 0.1%. Yet this decrease can be attributed almost exclusively to the domain Nonhispanic White. With alternative (6) based on MER data and probabilities, there were significant increases in the estimates of every domain. The relative differences under alternative (6) are larger in magnitude than for earlier alternatives, but all have an absolute magnitude of less than 0.4%. There are several relative differences greater than 0.3% in absolute value: for Nonhispanic Black, Hispanic, and American Indian not on Reservation.

6. Observations

The observations given here pertain to the third type of missing data, assigning probabilities to unresolved people in the A.C.E. It is important to note that the A.C.E. procedures were specified well before the conduct of the census and the A.C.E. The early deadlines were due to (1) the very tight schedule coordinating many separate but interrelated activities, and (2) the need for a process open to the scrutiny of policy makers as well as statistical experts. Although one can learn much about the missing data and the relevant correlation structures by examining the responses as they are collected, making decisions after seeing the data might have been construed as manipulating the results of an operation that had serious political implications.

In this light, one can look back and realize various ways to improve the process – too late to hange the procedures.

This does not imply that we did not react to information made available unexpectedly during the processing of the data. We knew that the post-match follow-up operation would help resolve some cases, especially those whose true residence on Census Day was uncertain. Much other information was collected in these interviews, but we did not anticipate seeing the details. However, due to an intensive keying of the follow-up interview forms at the Bureau's processing center, some additional information was made available during the missing data operation. At that time, we added several match code groups not originally in the plan: group 8 for resident status; 11 and 12 for enumeration status. Separating the people in these groups allowed us to assign probabilities that were quite different – and, we believe, more accurate – from what they would have received.

Different models, imputation cells, or data could have been used to assign probabilities for unresolved cases. The values determined through logistic regression were quite similar on average, and may or may not have had an effect on the resulting population estimates. In section 5 it was shown that ignoring some of the demographic variables would have made a difference in the match rate, but probably not in the rate of correct enumeration. Basing the probabilities on data collected in the Matching Error Study or the Measurement Error Reinterview (not yet available during the A.C.E.) could have made a larger difference still. But it is unclear which one might have made an improvement; using MES data would have lowered the population estimates, while using MER data would have increased them.

Weighing the various results, one is constantly reminded that, when assigning probabilities to people with unresolved status, match code group was the most important variable. It worked well for resident and enumeration status, but could not be effectively used for match status. The problem there – perhaps the biggest hole in our procedures – is once again that almost all of the unresolved matches, and over half of the unresolved residents, were people with insufficient information for matching. Little information was collected on these cases, and almost all of them were not sent through the matching process or follow-up. Further, almost none of these people were included in any post-A.C.E. evaluations. In future tests a concerted effort should be made to obtain real information about the status of such people.

Acknowledgements

The authors thank Eric Schindler and Doug Olson for computing dual-system estimates and their standard errors under alternative procedures; Tom Belin, UCLA, for making available imputation probabilities under logistic regression models; and Mary Frances Zelenak and Ha Nguyen for compiling summaries of the extent of missing data in the A.C.E. This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U. S. Census Bureau.

References

Anderson, M.J., and Fienberg, S.E. (1999). *Who Counts? The Politics of Census-Taking in Contemporary America*. New York: The Russell Sage Foundation.

Belin, T. (2001). Evaluation of unresolved enumeration status in 2000 Census Accuracy and Coverage Evaluation program. Unpublished report, prepared by Datametrics, Inc., for the U.S. Census Bureau.

Belin, T., Diffendal, G., Mack, S., Rubin, D., Schafer, J. and Zaslavsky A. (1993). Hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation. *Journal of the American Statistical Association*, 88, 1149-1166.

Cantwell, P.J. (2001). Accuracy and Coverage Evaluation Survey: Specifications for the missing data procedures. *DSSD Census 2000 Procedures and Operations Memorandum Series*, Q-62.

Cantwell, P.J., Mcgrath, D., Nguyen, N. and Zelenak, M.F. (2001). Accuracy and Coverage Evaluation: missing data results. *DSSD Census 2000 Procedures and Operations Memorandum Series*, B-7*.

Childers, D. (2000). The Design of the Census 2000 Accuracy and Coverage Evaluation. *DSSD Census 2000 Procedures and Operations Memorandum Series*, Chapter S-DT-1.

Fenstermaker, D. (2000). The Accuracy And Coverage Evaluation: sample design summary. *DSSD Census 2000 Procedures and Operations Memorandum Series*, R-33.

Haines, D. (2003). A.C.E. Revision II results: changes in estimated net undercount. *DSSD A.C.E. Revision II Memorandum Series*, PP-58

Hogan, H. (1993). The Post-Enumeration Survey: Operations and results. *Journal of American Statistical Association*, 88, 1047-1060.

Hogan, H. (2003). The Accuracy and Coverage Evaluation: Theory and design. *Survey Methodology*, 29, 129-138.

Ikeda, M., Kearney, A. and Petroni, R. (1998). Missing data procedures in the Census 2000 Dress Rehearsal Integrated Coverage Measurement sample. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 617-622.

Kearney, A., and Ikeda, M. (1999). Handling of missing data in the Census 2000 Dress Rehearsal Integrated coverage measurement sample. *Proceedings of the Survey Research Section*, American Statistical Association, 468-473.

Keathley, D., Kearney, A. and Bell, W. (2001). ESCAP II, Analysis of missing data alternatives for the Accuracy and Coverage Evaluation. Executive Steering Committee for A.C.E. Policy II (ESCAP II) Report 12.

Killion, R.A. (2000). Measurement Error Reinterview Sample Selection. *Planning, Research, and Evaluation Division TXE/2010 Memorandum Series*, CM-MER-S-01.

Kostanich, D. (2001). Accuracy and Coverage Evaluation Survey: computer specifications for Person Dual System Estimation (U.S.) - Re-issue of Q-29. *DSSD Census 2000 Procedures and Operations Memorandum Series*, Q-37.

Navarro, A., and Olson, D. (2001). Accuracy and Coverage Evaluation: effect of targeted extended search. *DSSD Census 2000 Procedures and Operations Memorandum Series*, B-18*.

Peterson, C.G.J. (1896). The yearly immigration of young plaice into the Limfjord from the German Sea. *Report of the Danish Biological Station*, 6, 1-48.

Petroni, R. (1997). Effect of using the 1996 ICM characteristic imputation and probability modeling methodology on the 1995 ICM *P* and *E* – sample data. *DSSD Census 2000 Dress Rehearsal Memorandum Series*, A-20.

Petroni, R. (1998a). Effect of different methods for calculating match and residence probabilities for the 1995 *P* – sample data. *DSSD Census 2000 Dress Rehearsal Memorandum Series*, A-23.

Petroni, R. (1998b). Effect of different methods for calculating correct enumeration probabilities for the 1995 *E* – sample data. *DSSD Census 2000 Dress Rehearsal Memorandum Series*, A-28.

Petroni, R. (1998c). Effect of using simple ratio methods to calculate *P* – sample residence probabilities and *E* – sample correct enumeration probabilities for the 1995 data. *DSSD Census 2000 Dress Rehearsal Memorandum Series*, A-30.

Petroni, R. (2001). EFU Sample Design, Stratification, Selection, and Weighting. Planning, Research, and Evaluation Division *TXE/2010 Memorandum Series*, CM-GES-S-02-R2.

Rubin, D.B. (1976). Inference and Missing Data. *Biometrika*, 63, 581-592.

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