Analyzing the Imputations for Nonresponse in the 2020 Census Post-Enumeration Survey[1]

Michael Beaghen, Richard N. Turner
U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233

Abstract

The Post-Enumeration Survey (PES) estimated the net coverage error of the 2020 Census. Like all surveys, the PES had to deal with missing or incomplete response data. Some respondents did not answer specific questions needed to estimate the population size. When this happened, we imputed values to fill in the missing data. This paper describes the methodology and results of the processes the PES undertook to treat missing data.

Keywords: Census-coverage error, logistic regression

1. Introduction

This paper provides an overview of how missing data were handled in the 2020 Post-Enumeration Survey (PES) for the United States person estimates. It describes the missing data procedures used to support the estimation of net coverage error. It focuses on the imputation of match and correct enumeration statuses used in the dual-system estimation.

This paper summarizes a larger report on missing data for people in the PES ((Beaghen et al., 2022). Documentation of the overall PES design can be found in Kennel (2019). Documentation of the estimation methodology can be found in Zamora (2022). Discussion of the source and the accuracy of the person estimates can be found in Marra and Kennel (2022).

The paper is organized as follows. Section 2 starts with background on the 2020 PES and why there was missing data in the 2020 PES. Section 3 discusses the important concept of sufficient information for dual-system estimation. Section 4 motivates the use of logistic regression models to assign missing values. Section 5 describes the logistic regression models used to impute for correct enumeration status for E-sample people with missing statuses. Lastly, Section 6 gives a less detailed discussion on the imputation for missing P-sample statuses.

[1] Any views expressed are those of the authors and not those of the U.S. Census Bureau. The Census Bureau's Disclosure Review Board and Disclosure Avoidance Officers have reviewed this data product for unauthorized disclosure of confidential information and have approved the disclosure avoidance practices applied to this release. CBDRB-FY22-137 and CBDRB-FY22-244.

2. Background on the 2020 Post-Enumeration Survey

The Census Bureau conducted a Post-Enumeration Survey (PES) to assess the 2020 Census' coverage of population and housing units. Census person coverage errors included omissions, duplication, people enumerated in the wrong place, and enumeration of people who should not have been enumerated. The final 2020 PES sample size was roughly 160,000 housing units across the 50 states and the District of Columbia (Marra and Kennel, 2022). The 2020 PES methodology required an independent listing of housing units and people in housing units in a sample of geographies. This independence from the Census was necessary to satisfy the requirements for the dual-system estimator. The PES listings were matched to the census listings of housing units and people. Accurate matching included automated probability-based matching, followed by clerical matching, and by field interviewing to resolve differences in the listings.

The PES used dual-system estimation to estimate the population size of the nation. By comparing the PES population estimate to the 2020 Census, the PES estimated the net coverage error of the 2020 Census count of people. The dual-system estimator used by the 2020 PES required both a probability of match and a probability of correct enumeration; refer to Zamora (2022) for details on how the dual-system estimates were calculated.

The PES consisted of two samples: a sample of the population or **P sample**, and a sample of census enumerations or **E sample**. The P sample of housing units and people in housing units was enumerated independently of the census. The E sample consisted of census housing units and census person enumerations in housing units in the same sample areas as the P sample. We used the P sample to estimate the match probability and the E sample to estimate the correct enumeration probability.

When clerical matching and field interviewing did not provide enough information to determine the match status of P-sample or the correct enumeration status of E sample people, these statuses were imputed with logistic regression models. This paper documents the methodology and the results of these imputations for statuses.

2.1 Post-Enumeration Survey Missing Data

Before calculating dual-system estimates, we accounted for missing data in the P and E samples. We encountered three types of missing data.

1. *Household-level noninterviews in the person P sample.* For some of these noninterviews, the household could not be contacted or the interview was refused or not completed. However, more commonly, the information provided by the respondent was not complete enough to accurately match anyone in the household to the census. The noninterview adjustment spread the weights of household noninterviews to similar households that were interviewed. Refer to Beaghen et al. (2022) for details. We do not discuss the noninterview adjustment further in this paper.

2. *Unresolved statuses in the P and E samples.* When we refer to status, we usually mean the answer to a question we needed to estimate coverage. There are four statuses discussed in this memo:
   - E-sample enumeration status: Was a person correctly enumerated or erroneously enumerated?
   - P-sample inclusion status: Did a person meet the requirements for being in-scope for the PES?
   - P-sample mover status: Did a person move between April 1, 2020 and the PES interview?
   - P-sample match status: Was a person in the PES correctly counted in the 2020 Census?

   The statuses provided the information needed to calculate dual-system estimates. Missing statuses arose when we did not have enough information about a person to make a confident determination. When a status was missing, we imputed a probability for that status using information available about the person and about resolved cases with similar characteristics.

3. *Missing demographic characteristics in the P sample*. This situation occurred when a person was missing age, sex, relationship, tenure, race, or Hispanic origin. The characteristic imputation methods are discussed in Phan and Lawrence (2022). We do not discuss them in this paper.

2.2 Why was there Missing Data in the Post-Enumeration Survey?

The person statuses provided the information needed to calculate dual-system estimates. Missing statuses arose when we did not have enough information about a person to make a confident determination. When a status was missing, we imputed a probability for that status using information available about the person and from resolved cases with similar characteristics.

Missing data in the PES resulted from failure to obtain all needed information from interviews. The interviews which determined the person statuses were the PES Person Interview and the PES Person Followup interview (refer to Kennel, 2019, for details of the PES design).

If neither the PES Person Interview nor the Person Followup interview provided the information needed to determine one or more of a P-sample person's statuses, then those statuses were considered missing. Note that if the Person Interview failed to collect sufficient information for a person, then it either was missing all its statuses or it was processed in the Noninterview Adjustment (refer to Beaghen et al., 2022).

The status of an E-sample person enumeration was usually determined in one of two ways. If it matched (was determined to represent the same individual) to a P-sample person, it took the Census Day status of that matching P-sample person. Thus if the matching P-sample person was a valid person on Census Day, the E-sample person was assigned a correct enumeration status. If the matching P-sample person was not a valid Census Day person, then the E-sample person had a status of erroneous enumeration. If the matching P sample person was missing its status then the E-sample person also had a missing correct enumeration status. The second way the E-sample person enumeration was determine was from the Person Followup interview. If this interview failed to provide the information

needed to determine the E-sample person's correct enumeration status, then that status was considered missing.

As we point out in the later sections, there were more P- and E-sample people with missing statuses in the 2020 PES than in the 2010 PES. The 2020 PES was conducted during the COVID-19 pandemic. COVID-19 may have made some people reluctant to respond to interviewers, or pressured those willing to respond to keep interviews short and possibly less thorough (Khubba and Marra, 2022). Also, much of the PES interviewing was delayed, and respondents might not have been able to remember the required information.

3. Preliminaries: Sufficient Information for Dual-System Estimation

The concept of sufficient information for dual-system estimation is important throughout this document and defined here. Person records with sufficient information for dual-system estimation had adequate information to uniquely identify an individual. Person records with insufficient information for dual-system estimation did not meet the minimum threshold to uniquely identify a person. We could not determine with confidence the inclusion, match, or enumeration statuses of insufficient information cases using the PES matching and field operations. For many of the insufficient information cases, a name was missing. For simplicity, we use the terms "sufficient information" and "insufficient information" throughout the remainder of this document.

Table 1 and Table 2 show for the P and E samples, respectively, the percentage of insufficient information cases in the 2020 PES and the 2010 CCM[2]. We note higher rates of insufficient information for both the P and E samples in the 2020 PES.

Table 1: P-Sample Insufficient Information Counts

	2020		2010	
	Total	Insufficient Information Count	Total	Insufficient Information Count
All PES Listed Cases	345,000	51,000	393,000	13,000
Post-Noninterview Adjustment P-sample Cases	301,000	12,500	383,000	6,400

Source: U.S. Census Bureau, Decennial Statistical Studies Division, 2020 Post-Enumeration Survey (May 2022 release) and 2010 Census Coverage Measurement Survey.

Through interviews that were independent of the 2020 Census, the PES listed 345,000 people. Of those people, 51,000 did not have sufficient information for dual-system estimation. If everyone in the household had insufficient information, the household was called a noninterview and its weight distributed among responding households in a process called the Noninterview Adjustment. After the Noninterview Adjustment, there were still 12,500 people who had insufficient information, but they were in households where at least one person had sufficient information. These had their inclusion and match statuses imputed. Note that there were also person listings with sufficient information whose

[2] The post-enumeration survey in 2010 was called the Census Coverage Measurement survey.

weights were distributed in the Noninterview Adjustment process, or were removed from PES processing because of PES data editing rules.

We treated insufficient information E-sample cases as erroneous for the estimation of net coverage error. There were about 40,000 such cases in the 2020 PES, as opposed to just 13,000 in the 2010 PES.

Table 2: E Sample Insufficient Information Counts

2020		2010	
Total	Insufficient Information Count	Total	Insufficient Information Count
397,000	40,000	384,000	13,000

Source: U.S. Census Bureau, Decennial Statistical Studies Division, 2020 Post-Enumeration Survey (May 2022 release) and 2010 Census Coverage Measurement Survey.

4. Logistic Regression Models for Status Imputations for Net Coverage Error

After finishing all data collection activities, there remained E-sample enumerations without enough information to determine the correct enumeration status, and P-sample people without enough information to determine the inclusion or match statuses. A common reason for unresolved E-sample and P-sample statuses was the lack of reported data from the PES interviews needed to determine the correct enumeration, inclusion, or match status. The 2010 and 2020 PES imputed values for the missing statuses using survey-weighted logistic regression models fit on the resolved data.

The statuses we imputed were binary, thus all logistic regression models described in this report had binary dependent variables. Resolved cases representing the "yes" category (i.e., included in the P sample, match, or correct enumeration) were assigned a value of 1 for the dependent variable. Resolved cases representing the "no" category (i.e., not included in the P sample, nonmatch, or erroneous enumeration) were assigned a value of 0 for the dependent variable.

Missing statuses were imputed with a fraction between 0 and 1. For example, a P-sample case might have a predicted probability of 0.80 for inclusion status, meaning 80 percent of the record's weight was counted as being in the P sample and 20 percent was not. If we also imputed the match status, some portion of the 80 percent contributed to a match and the reminder was a nonmatch. Thus, 50 percent of the record's weight might be a match, 30 percent a nonmatch, and 20 percent out-of-scope.

4.1 Example of a Covariate: Before Followup Match Code Group

The goal of imputation with logistic regression modeling is to use existing information to make a prediction about the missing statuses. For example, certain demographic covariates have a history of being correlated with both erroneous enumerations and nonmatches. These include owner/renter, age, sex, race, and Hispanic origin. Including these covariates in a logistic regression model yielded better predictions of status than naively substituting an overall mean probability.

A covariate that, by itself, had noticeable predictive power was the Before Followup Match Code Group (BFUMCG). This variable existed in different forms for both the E sample

and the P sample, though we only discuss the E-sample variable here. Table 3 shows the distribution of correct and erroneous enumerations by the values of BFUMCG. During clerical matching every census record was reviewed, and staff determined if a follow-up interview was needed to get more information about the person's enumeration status. The Before Followup Match Code Group summarized why a follow-up interview was or was not necessary. We see that Resolved Before Followup had 229,000 people resolved as correct enumerations and 6,100 resolved as erroneous enumerations; but only a relatively small number of people had an unresolved status, 750. Most people needing imputation were in the Whole Household Nonmatch and Unclassified Inclusion Status of Matching P-sample. Because the correct enumeration rates of the resolved cases differed by the Before Followup Match Code Group and the share of the unresolved cases differed by these groups as well, this variable (or a suitable version of it) was used in nearly every imputation model.

Table 3: Counts of Correct and Erroneous Enumerations by Before Followup Match Code Group

Before Followup Match Code Group	Correct Enumerations	Erroneous Enumerations	Resolved Correct Enumeration (Percent)	Number of Unresolved
Resolved Before Followup	229,000	6,100	97.4	750
Possible Matches	850	30	96.6	150
Conflicting Household	4,800	500	90.6	3,000
Partial Household Nonmatch	16,500	2,200	88.2	4,800
Whole Household Nonmatch	33,000	4,100	88.9	15,000
Duplicate	2,500	650	79.4	700
Unclassified Inclusion Status of Matching P-sample	10,000	300	97.1	21,500
Insufficient Information	0	40,000	0.0	0
Total	297,000	54,000	84.6	46,000

Note: Counts may not sum to totals shown because of rounding.
Source: U.S. Census Bureau, Decennial Statistical Studies Division, 2020 Post-Enumeration Survey (May 2022 release).

One can better understand the predictive power of the covariate BFUMCG with some knowledge of the PES processing. E-sample enumerations that matched to a valid and nonmover P-sample person were coded as a correct enumeration and not sent to follow-up (they were BFUMCG Resolved Before Followup) because we already received an independent verification of the enumeration from the Person Interview field operation. That is, the E-sample match to an in-scope P-sample person indicated that the E-sample enumeration also represented an in-scope person or correct enumeration. Erroneous enumerations, on the other hand, often could not be matched to valid P-sample people. Thus the E-sample enumerations that did not match to an in-scope P-sample person had higher probabilities of being erroneous. These nonmatching E-sample cases included the BFUMCG groups Conflicting Household, Partial Household Nonmatch, and Whole Household Nonmatch.

Note that most of the approximately 21,500 E-sample enumerations in the Before Followup Match Code Group labeled Unclassified Inclusion Status of Matching P-sample initially had an enumeration status assigned, but were blanked out in response to concerns about their initial status assignment. These enumerations were matched to P-sample person listings for which we had insufficient information. The PES should have sent them to a follow-up interview to determine the correct enumeration status of the E-sample enumeration, but failed to do so. Since we were uncertain that the P-sample person was a valid person, we could not assume a matching E-sample enumeration was a correct enumeration. So we imputed their enumeration status.

The Before Followup Match Code Group effectively partitioned the resolved cases into cells with different correct enumeration rates. For example, the category Resolved Before Followup had a very high correct enumeration rate, 97.4 percent, while the partial and whole household nonmatch groups had lower correct enumeration rates, 88.2 percent and 88.9 percent, respectively. This kind of partitioning of the data into groups with similar correct enumeration rates within the group, but with differing rates between the groups, is a key characteristic of a powerful covariate for imputation.

To assess the usefulness of a covariate it is also important to consider the distribution of the unresolved cases. Categories that have high numbers of resolved cases but few unresolved cases can yield a logistic regression model with a high-level of fit, and yet be of minimal predictive value in practice. We see this with the two groups with the largest number of resolved cases, Resolved Before Followup and Insufficient Information. They had only 750 and 0 unresolved cases each, respectively. However, some of the other groups demonstrate the potential utility of Before Followup Match Code Group. For example, there were about 15,000 unresolved cases with Whole Household Nonmatch, with a resolved correct enumeration rate of 88.9 percent, and about 21,500 unresolved cases with Unclassified Inclusion Status of Matching P-sample People, with a resolved correct enumeration rate of 97.1 percent.

5. Imputation of E-Sample Correct Enumeration Status for Net Coverage Error

To calculate the dual-system estimates, we needed to assign an enumeration status to each E-sample person enumeration. We defined an E-sample enumeration as a correct enumeration for the estimation of net coverage error if it was an enumeration with sufficient information that corresponded to a person who should have been counted in the block search area[3] in a housing unit on Census Day.

Enumerations not meeting these criteria were erroneous enumerations. Erroneous enumerations included people who were born after Census Day or who died before Census Day, fictitious enumerations, and people counted in the wrong location (i.e., people who should have been counted somewhere outside of the block search area). In addition, if two or enumerations referred to the same person, one was called correct and the others erroneous due to duplication.

Note that E-sample enumerations with insufficient information were treated as erroneous enumerations for net coverage error estimation. We did this because we could not match

[3] To be more precise, it was not the 'block search area' but the 'basic collection unit search area.' A basic collection unit was the smallest geographic level for 2020 Census data collection and roughly corresponded to a block. Refer to Hogan (2003) for more details on the search area.

P-sample people to them accurately. Some of the P-sample records who represented the same person as an insufficient information census enumeration could be matched. But other P-sample records who represented the same person as an insufficient information census enumeration would not be matched and would yield false nonmatches. False nonmatches would bias the match rate and the DSE used to calculate the census net coverage error. We avoided introducing this bias in the DSE by treating all E-sample insufficient information cases as erroneous enumerations and all P-sample people who matched to insufficient information census enumerations as nonmatches.

Table 4 has a summary of the enumeration status for E-sample people. The rate of missing enumeration status in the E sample was higher for the 2020 PES than the 2010 CCM. Part of the increased rate resulted from the special fix of E-sample people that matched to P-sample people with insufficient information person statuses who did not go to follow-up (refer to Beaghen et al., 2022). However, even without this fix the E-sample unresolved rate would have been noticeably higher than the 2010 CCM rate. There were several factors that may have contributed to the higher unresolved rates in the 2020 PES. They included difficulty conducting interviews because of COVID-19, the greater amount of missing characteristics of E-sample enumerations that made matching and follow-up more difficult, and more insufficient information cases in the P sample, which would have required more E-sample enumerations going to a follow-up interview.

Table 4: 2020 PES and 2010 CCM Person Enumeration Status

	2020 PES	2010 CCM
Total E-sample Enumerations	397,000	384,000
Number of Resolved Enumerations	351,000	365,000
Number of Unresolved Enumerations	46,000	18,500
Unresolved Enumeration (Percent)	11.6	4.8

Note: Counts may not sum to totals shown because of rounding.
Source: U.S. Census Bureau, Decennial Statistical Studies Division, 2020 Post-Enumeration Survey (May 2022 release) and 2010 Census Coverage Measurement Survey.

5.1 Logistic Regression Modeling to Impute for Missing Correct Enumeration Status

We imputed the probability of correct enumeration using logistic regression for cases with a missing enumeration status. We used one of two models to predict the correct enumeration probability for unresolved E-sample enumerations. The models were fit on the resolved cases (there were 351,000 resolved cases in the 2020 PES; refer to Table 4). Each used the same core set of main effects, which included demographic characteristics such as age and sex as well as a proxy interview flag. A full listing of the variables used in each model can be found in Table A1 of the Attachment. A description of each variable and its categories can be found in Table A2 of the Attachment. The weight used in the models was the original E-sample weight (the inverse of the probability of selection).

The first model included nine additional indicator variables, while the second model excluded these variables. One of these variables was a duplicate link flag that indicated whether the E-sample person was linked to another census enumeration as a possible duplicate. The other eight flag variables indicated whether a person had certain types of

additional addresses attached to them. These additional address flags included outmover, seasonal, inmover, college, relative, military, job, and group quarters address flags.

We made this distinction with the existence of an address flag because we had a clear response with a "yes" response. In contrast, if a person record did not have any additional addresses attached to them, it was not clear if the respondent did not have an alternative address or was not responding to the question. The respondent might have neglected to provide this additional information or might not have known this information for the people they were responding for. We did not want the predicted probability for people without the address flags to be influenced by the effects of the address flags in the model.

5.2 Results for Imputing the Correct Enumeration Status

Table 5 shows the overall effect of imputation on the correct enumeration rate. While the difference could appear small at first glance, at the national level it would have a noticeable impact. Note that the overall correct enumeration rate presented in Table 5 differs from estimates of the component "correctly enumerated in the BCU search area," as presented in Table 2 of Khubba et al. (2022). The component of coverage estimates included imputations for E-sample insufficient information cases, whereas the estimates of net coverage error treated the E-sample insufficient information as erroneous. Imputation increased the correct enumeration rate from 86.75 percent to 87.16 percent. This increase in the correct enumeration rate resulted in an increase in the dual-system estimate of the population size.

Table 5: Correct Enumeration Rate With and Without Imputation for the 2020 PES (Weighted)

	Resolved Cases Only	After Imputation
Correct Enumeration Rate in Percent	86.75	87.16

Source: U.S. Census Bureau, Decennial Statistical Studies Division, 2020 Post-Enumeration Survey (May 2022 release).

Of course, the imputed correct enumeration rates varied among the unresolved cases with different characteristics. Table 6 shows the 25th percentile, the median, and the 75th percentile of the imputed values. Over 75 percent of the unresolved cases were imputed with a relatively high correct enumeration rate over 0.8761.

Table 6: Distribution of Imputed Correct Enumeration Probabilities for the 2020 PES (Unweighted)

25th Percentile	Median	75th Percentile
0.8761	0.9427	0.9785

Source: U.S. Census Bureau, Decennial Statistical Studies Division, 2020 Post-Enumeration Survey (May 2022 release).

6. Imputation of P-Sample Inclusion, Mover, and Match Statuses for Net Coverage Error

Dual-system estimation required us to determine whether each person in the P sample matched to an enumeration in the Census[4]. After all PES data collection activities were completed, there remained people listed in the P sample without enough information to determine an inclusion, mover, or match status. This section provides an overview of issues pertaining to missing statuses in the P sample.

6.1 The P-Sample Statuses

There were four P-sample statuses relevant to dual-system estimation: inclusion, mover, inmover match, and nonmover match.

- Inclusion status: Whether a person listed during the PES Person Interview should have been included in the P sample.
- Mover status: Whether a person was a nonmover—i.e., lived at the PES sample address on both Interview Day and Census Day—or whether they were an inmover—i.e., moved into the PES address after Census Day.
- Inmover match status: Given that a person was an inmover, whether they matched to a census enumeration at their Census Day address.
- Nonmover match status: Given that a person was a nonmover, whether they matched to a census enumeration at the PES sample address.

First we discuss the inclusion status. Before determining the match status, we needed to determine which PES person listings were in-scope for the P sample. People living in Group Quarters facilities (for example a prison, college dorm, or nursing home) and Remote Alaska areas on Census Day, and visitors are examples of people who were not eligible to be in the P sample. Sometimes we did not have enough information to determine if someone should have been included in the P sample. Thus, we imputed the inclusion status for such people.

The PES Person Interview included questions about where everyone in the household on the PES Interview day was living on Census Day. If we did not have enough information to determine if the person was in-scope for the PES, we imputed the P-sample inclusion status.

Once the inclusion status was determined for all people listed in the PES Person Interview, we had to determine their mover status. It was possible that the person moved into the PES housing unit between Census Day and the PES Interview Day, in which case they were an inmover. If the person lived at the same address on both Census Day and Interview Day, they were a nonmover. The PES Person Interview asked questions about where people were living around Census Day. The reported Census Day addresses were compared to the Interview Day address by clerical matching staff. If the clerical matching staff were not able to determine if a person was an inmover or nonmover, we imputed their mover status.

Mover status was important because it determined the search area for matching the PES to the Census (refer to the next paragraph for details). In past post-enumeration surveys, mobility has been a major factor in our ability to determine match status. Inmovers were

[4] Refer to Zamora (2022) for details on how the dual-system estimates were calculated.

generally more difficult to match to the census than nonmovers and had higher unresolved match rates. For this reason, we imputed match status separately for inmovers and nonmovers. And, because we did not always know whether a person was an inmover or nonmover, we had to impute mover status before imputing match status.

To limit the error of false matches (calling a P-sample person and census enumeration a match when they referred to different people), the matching was done in a limited search area. The search area for an inmover consisted of the block containing the address they reported being at on Census Day and the ring of surrounding blocks. The search area for a nonmover consisted of the block containing the PES sample address and the ring of surrounding blocks. A P-sample person was considered a match only if they matched to a census enumeration in the correct search area. If they matched to an enumeration outside the search area, they were classified as a nonmatch for dual-system estimation. For more information on the PES search area, refer to Hogan (2003).

Thus, the P-sample people required up to four separate imputations. The following equation indicates how information on the statuses was combined to calculate the overall probability that a P-sample person matched to a census enumeration.

$$p_{match,j} = p_{inmover,j} \times p_{match|inmover,j} + \left(1 - p_{inmover,j}\right) \times p_{match|nonmover,j}$$

where for person record j,

- $p_{match,j}$ is the overall probability of being a match.
- $p_{inmover,j}$ is the probability of being an inmover.
- $p_{match|inmover,j}$ is the inmover match probability.
- $p_{match|nonmover,j}$ is the nonmover match probability.

6.2 Unresolved P-Sample Statuses

A person for whom we determined a given status is referred to as "resolved" for that status. For instance, people with a resolved inclusion status were those that were identified either as in the P sample or as not in the P sample. It was not always possible to determine the inclusion, mover, inmover match, or nonmover match status of a person listed during the Person Interview—rendering them "unresolved" for that status or those statuses. It was possible that a P-sample person could be resolved for one or more status but not for others.

P-sample people with at least one unresolved status fell into one of two categories:

- Sufficient information for dual-system estimation.
- Insufficient information for dual-system estimation.

Refer to Section 3.0 for the definitions of sufficient and insufficient information for dual-system estimation.

Table 7 (shown at the end of the paper) presents counts and rates of P-sample cases that were missing each status. The first row, "P-Sample Inclusion Status," shows the raw counts. However, the counts for mover status and the match statuses were multiplied by the person's probability of being in the P sample (imputed to be greater than 0 but less than 1 for those with an unresolved inclusion status). For example, consider how we obtained the mover status counts. Of the records with a resolved inclusion status, around 262,000

were known to be in the P sample (refer to Table 14 in Beaghen et al.) and thus each counted as one record in the mover status calculations. These records were summed with the roughly 21,000 records with an unresolved inclusion status, multiplied by their respective predicted inclusion probabilities. The predicted inclusion probabilities equaled approximately 0.80 on average, leading to the Mover Status total of 279,000. The match status counts additionally account for the probability of being an inmover (imputed between 0 and 1 for those with an unresolved mover status).

The results show that the PES had to rely on imputation procedures to a greater degree than the 2010 CCM. Indeed, the overall unresolved inclusion rate was over twice as large in 2020 as 2010 (6.98 percent vs. 2.87 percent) as was the unresolved match rate (5.02 percent vs. 1.90 percent)[5].

6.3 Logistic Regression Modeling to Impute for Missing P-Sample Statuses

As with the imputation for correct enumeration status (Section 5), we used survey-weighted logistic regression models to impute missing P-sample statuses. For fitting the inclusion status models, the weight used was the product of the sampling weight and the noninterview adjustment factor. For the mover and match models, the weight used was the product of the sampling weight, the noninterview adjustment factor, and the final inclusion probability[6]. For details refer to Beaghen et al. (2022).

Table 8 presents the weighted match rate before and after the status imputation processes. The "Resolved Match Rate" is limited to people for whom inclusion status, mover status, and match status could be directly measured because they were not missing any data required for determining these statuses. The "After Imputation" is the match rate including unresolved cases after all imputation was finished. As we see, the imputation decreased the match rate from 86.77 percent to 84.98 percent. This decrease in the match rate increased the dual-system estimate of the population size.

Table 8: Match Rate With and Without Imputation for the 2020 PES (Weighted)

	Resolved Cases Only	After Imputation
Match Rate (Percent)	86.77	84.98

Source: U.S. Census Bureau, Decennial Statistical Studies Division, 2020 Post-Enumeration Survey (May 2022 release).

[5] These percentages account for differences in estimated inclusion and mover probabilities. Refer to Table C1 in Beaghen et al. (2022) for the raw numbers of cases with unresolved statuses.

[6] Note that this logistic regression-based approach to mover status imputation departed from the procedure used in the 2010 CCM. In 2010, a cell mean methodology was used to impute mover status probabilities for unresolved cases, where the cells were based on the BFU match group variable for sufficient information cases and on the BFU insufficient information group variable for insufficient information cases.

Table 7: Unresolved Rates for Inclusion, Mover, and Match Statuses (Unweighted)

	2020				2010			
	Total	Resolved	Unresolved	Unresolved (Percent)	Total	Resolved	Unresolved	Unresolved (Percent)
P-Sample Inclusion Status	301,000	279,000	21,000	6.98	383,000	372,000	11,000	2.87
P-Sample Mover Status	279,000	264,000	15,500	5.56	352,000	345,000	7,600	2.16
P-Sample Total Match Status	279,000	265,000	14,000	5.02	352,000	346,000	6,700	1.90
P-Sample Inmover Match Status	22,000	16,500	5,400	24.55	28,000	24,500	3,300	11.79
P-Sample Nonmover Match Status	257,000	249,000	8,600	3.35	324,000	321,000	3,300	1.02

Notes:

1. Counts for the mover status row are multiplied by the probability of inclusion.
2. Counts for the inmover match status row are multiplied by the probability of inclusion and the probability of being an inmover.
3. Counts for the nonmover match status row are multiplied by the probability of inclusion and the probability of being a nonmover.
4. Counts for total match status were calculated by summing the respective counts in the inmover match status and nonmover match status rows. This approach is valid because the inmover match status and nonmover match status figures are multiplied by, respectively, the inmover and nonmover probabilities.
5. Counts may not sum to totals shown because of rounding.

Source: U.S. Census Bureau, Decennial Statistical Studies Division, 2020 Post-Enumeration Survey (May 2022 release) and 2010 Census Coverage Measurement Survey.

References

Beaghen, M., Turner, R., Jost, M., and Marra, E., "2020 Post-Enumeration Survey Estimation Methods and Results: Missing Data for Person Estimates," 2020 Post-Enumeration Survey Memorandum series #2020-J-05, U.S. Census Bureau, 2022.

Hogan, H., "The Accuracy and Coverage Evaluation: Theory and Design," Survey Methodology, 29, U.S. Census Bureau, 2003, pp. 129–138.

Kennel, T., "The Design of the Post-Enumeration Survey for the 2020 Census," DSSD 2020 Post-Enumeration Survey Memorandum Series #2020-B-01, U.S. Census Bureau, 2019.

Khubba S., J. Hong, and K. Heim, "2020 Post-Enumeration Survey Estimation Report: Summary of Estimates of Coverage for People in the United States," DSSD 2020 Census Coverage Measurement Memorandum Series #2020 G-01, U.S. Census Bureau, 2022.

Khubba S., and E. Marra, "Covid-19 Impacts on the 2020 Post-Enumeration Survey," Proceedings of the 2022 Joint Statistical Meetings, 2022.

Marra, E., and T. Kennel, "Source and Accuracy of the 2020 Post-Enumeration Survey Person Estimates," DSSD 2020 Census Coverage Measurement Memorandum Series #2020 J-01, U.S. Census Bureau, 2022.

Phan, N., and J. Lawrence, "2020 Census Coverage Measurement Estimation Report: Characteristic Imputation Methods and Results," DSSD 2020 Post-Enumeration Survey Memorandum Series #2020-J-04, U.S. Census Bureau, 2022.

Zamora, J., "2020 Post-Enumeration Survey Estimation Design," DSSD 2020 Post-Enumeration Survey Memorandum series #2020-J-03, U.S. Census Bureau, 2022.

ATTACHMENT: Covariates for Correct Enumeration Status Imputation Model

Table A1: Model Variables Used in Status Imputation for Correct Enumeration Status

Variable	Correct Enumeration Model 1	Correct Enumeration Model 2
Race/Hispanic Origin Domain	X	X
Tenure	X	X
Sex	X	X
Age and Sex Group	X	X
Census Proxy Flag	X	X
Type of Census Response	X	X
Characteristic Imputation Flag	X	X
Relationship Type	X	X
BFU Match Code Group	X	X
Duplicate Link Flag	X	
Seasonal Address Flag	X	
Outmover Address Flag	X	
Inmover Address Flag	X	
Job Address Flag	X	
Military Address Flag	X	
Group Quarters Address Flag	X	
Relatives Address Flag	X	
College Address Flag	X	
Household with a Spousal Relationship	X	X
2010 CCM Correct Enumeration Rate by Tract	X	X
Relationship by Census Response Type Interaction	X	X
Relationship by Duplicate Link Flag Interaction	X	
Seasonal Address Flag by Duplicate Link Flag Interaction	X	

ATTACHMENT: Covariates for Correct Enumeration Status Imputation Model

Table A2: Model Variable Descriptions

Variable Description	Valid Values
Race/Hispanic Origin Domains	American Indian/Alaska Native On Reservations American Indian/Alaska Native Off Reservations Hispanic Non-Hispanic Black Native Hawaiian or Pacific Islander Asian White or Some Other Race
Tenure	Owner Renter
Age/Sex Group	0-4 5-9 10-17 18-24 Male 18-24 Female 25-29 Male 25-29 Female 30-49 Male 30-49 Female 50-64 Male 50-64 Female 65+ Male 65+ Female
Relationship Type	Nuclear Family Member Adult Child of the Householder Other Member of the Household
2010 CCM Tract-Level Person Match Rate	[Continuous Variable]
Spousal Household	No Spouse in Household Spouse Present in Household
Census Proxy Flag	Blank Household member on April 1 or Household member moved in after April 1 Other (multiple respondent types) Neighbor or other proxy respondent
CCM Correct Enumeration Rate by Tract	Continuous

ATTACHMENT: Covariates for Correct Enumeration Status Imputation Model

Variable Description	Valid Values
Type of Census Response (Session Context Code)	Internet Self-Response Paper Questionnaire Self-Response Electronic Enumeration or Paper Enumeration Administrative Records Coverage Followup
E-sample BFU Match Code Group	Resolved Before Followup Possible Matches Conflicting Household Partial Household Nonmatch Whole Household Nonmatch Unresolved Inclusion Status Duplicate Insufficient Information for Dual System Estimation
Duplicate Link Flag	No duplicate link attached to person Duplicate link attached to person
Seasonal Address Flag	No seasonal address attached to person Seasonal address attached to person
Outmover Address Flag	No outmover address attached to person Outmover address attached to person
Inmover Address Flag	No inmover address attached to person Inmover address attached to person
Job Address Flag	No Job address attached to person Job address attached to person
Military Address Flag	No Military address attached to person Military address attached to person
Group Quarters Address Flag	No Group Quarters address attached to person Group Quarters address attached to person
Relative Address Flag	No Relative address attached to person Relative address attached to person
College Address Flag	No College address attached to person College address attached to person