DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 2, 2017, pp. 455–475, http://dx.doi.org/10.1515/JOS-2017-0022

Comparison of 2010 Census Nonresponse Follow-Up Proxy Responses with Administrative Records Using Census Coverage Measurement Results

Mary H. Mulry[1] *and Andrew D. Keller*[1]

The U.S. Census Bureau is currently conducting research on ways to use administrative records to reduce the cost and improve the quality of the 2020 Census Nonresponse Followup (NRFU) at addresses that do not self-respond electronically or by mail. Previously, when a NRFU enumerator was unable to contact residents at an address, he/she found a knowledgeable person, such as a neighbor or apartment manager, who could provide the census information for the residents. This was called a proxy response. The Census Bureau's recent advances in merging federal and third-party databases raise the question: Are proxy responses for NRFU addresses more accurate than the administrative records available for the housing unit? Our study attempts to answer this question by comparing the quality of proxy responses and the administrative records for those housing units in the same timeframe using the results of 2010 Census Coverage Measurement (CCM) Program. The assessment of the quality of the proxy responses and the administrative records in the CCM sample of block clusters takes advantage of the extensive fieldwork, processing, and clerical matching conducted for the CCM.

Key words: 2020 Census; correct enumeration.

1. Introduction

The planning for the 2020 U.S. Census includes a program of research and testing aimed at developing methodology and processes to achieve cost containment and maintain quality. The program includes exploring and creating fundamental changes to the design, implementation, and management of the decennial census. A series of tests investigate proposed changes such as using adaptive strategies for conducting Nonresponse Followup (NRFU) of the housing units that do not self-respond in a census. The examined strategies include using administrative records and a variable number of contact attempts with the goal of reducing costs and improving data quality. One avenue of research focuses on whether administrative records can reduce the 2020 Census Nonresponse Followup (NRFU) fieldwork at addresses where the Census Bureau did not receive a self-response electronically or by mail. In previous censuses, when enumerators were unable to contact

[1] U.S. Census Bureau, Washington, DC 20233, 4600 Silver Hill Rd, Suitland, MD 20746, U.S.A. Emails: mary.h.mulry@census.gov and andrew.d.keller@census.gov

Acknowledgments: The authors thank Tom Mule for his useful advice and consultations. The authors thank Eric Slud and Richard Griffin, three anonymous referees, and the editors for their helpful comments on earlier versions of this manuscript. This report is released to inform interested parties and encourage discussion of work in progress. The views expressed on statistical, methodological, and operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

a household after a specified number of attempts, the instructions were to find a knowledgeable person. This person, perhaps a neighbor or apartment manager, who provided the census information for the residents, was called a proxy respondent. The question is whether a combination of federal and third-party databases provides better census information than the proxy responses.

Our study attempts to answer this question by comparing the quality of the proxy responses in the 2010 Census with administrative records for the same housing units. Previous studies have indicated differences in the quality of reporting of the population count and characteristics of the residents from household member respondents as opposed to proxy respondents. Both U.S. Constitutional and legislated uses of the census data involve the population counts and characteristics, such as age, sex, and race/Hispanic ethnicity so the collection of these data is fundamental to some government functions. Studies of proxy data following the 2000 Census found fewer missing characteristics in responses from household members versus proxies such as neighbors, postal workers or landlords (Chesnut 2005; Wolfgang et al. 2003). Regarding census coverage, Martin (1999) found that proxy reports of 'usual residence' increased undercoverage, particularly for unrelated household members. As part of research associated with the 2010 Census, King et al. (2012) found that self-report respondents provided more complete household membership than proxy respondents did.

The comparison of the quality of proxy responses and administrative records relies on the results of the 2010 Census Coverage Measurement (CCM) Program, which collected and processed the data used in forming estimates of census coverage error (Mule 2012). The goals of our study also include identifying variables that correlate with the quality of proxy responses and administrative records. Such variables, if they exist, would be useful in formulating decision rules for census processing. To provide context, our study also examines the quality of NRFU data from respondents who are household members and the administrative records available for the same addresses.

Ideally, one of the census tests could include a comparison of the proxy response for a housing unit and the administrative records for the same housing unit against a 'gold standard' interview conducted by a highly skilled interviewer with the residents of the housing unit. A determination could then be made as whether the proxy or the administrative records had better information, or whether they were of comparable quality. However, the 2020 Census testing cycle has a tight timeframe, which does not allow for a gold standard interview operation.

This article compares the quality of the 2010 Census NRFU housing units with proxy responses and the administrative records for the same housing units using the results of the 2010 Census Coverage Measurement (CCM) in a sample of block clusters. The approach is similar to a methodology discussed in Mulry and Spencer (2012). The administrative records files in our study come from two sources: (1) the Internal Revenue Service (IRS) 1040 forms filed in all months of 2010, and (2) the Medicare records for all months of 2010. The files from these two sources have the advantage of containing data for households.

This report describes the results of the first phase of our assessment. The second phase continues and includes a comparison of demographic characteristics of NRFU proxy responses and administrative records in corresponding housing units. Another aspect is to develop statistical models to identify the characteristics of NRFU housing units with

corresponding administrative records that have a high probability of being correct. The development of the models will consider characteristics of the households as well as geographic and socioeconomic variables available for census tracts and block groups from the U.S. Census Bureau's Planning Database (U.S. Census Bureau 2015). The Planning Database includes data from the U.S. Census Bureau's American Community Survey and the 2010 Census.

2. Research Approach

2.1. Research Questions

We aim to answer the following questions in order to produce information useful for the strategy design of contacting housing units during the 2020 Census NRFU:

- Are proxy responses for NRFU addresses more or less accurate than the administrative records available for the housing unit?
- What variables correlate with the accuracy of proxy responses for individual records and for records grouped by housing unit?
- What variables correlate with the accuracy of administrative records for individual records and for records grouped by housing unit?

2.2. Population

According to census residency rules, the correct address for a person's enumeration is his/her usual residence around Census Day, which is April 1 of the census year. The population under study is defined as the people whose Census Day residence is a housing unit enumerated in the 2010 Census NRFU by a proxy respondent, and administrative records are available for the housing unit. We consider the quality of two lists of the population using the criteria of whether the person is found at the correct location on Census Day according to census residency rules. One list of this population is the census enumerations, and the other list is the administrative records for the same housing units. For context, we also examine the quality of NRFU enumerations where the respondent is a household member and the administrative records at these addresses.

In this study, the definitions of the populations enumerated by proxy and household member respondents are operational and depend on the conduct of the 2010 Census operations. The housing units enumerated by household member respondents failed to self-respond by mail. The housing units enumerated by proxy failed to self-respond by mail, and none of the household members gave an interview to an NRFU enumerator. In 2010, enumerators had to make six contact attempts prior to taking a proxy interview. Therefore, our analyses, as well as the population definition, are conditional on the type of response observed in the 2010 Census. In addition, the analysis is conditional on the sources of administrative records that we consider.

2.3. Gold Standard

The assessment of the quality of the proxy responses and the records in the selected administrative files takes advantage of the extensive fieldwork, processing, and clerical

matching conducted for the CCM, which is the justification for using the CCM results as a gold standard. The 2010 CCM was designed to measure census coverage error with a post-enumeration survey composed of two samples, the population sample (P-sample) and the enumeration sample (E-sample). The former is a sample of housing units and persons selected independently of the census and designed to support the estimation of people missed in the census. Members of P-sample households are interviewed and then matched to the census on a case-by-case basis to determine whether they were enumerated in the census or missed. The E-sample is a sample of census enumerations (records) in the same areas as the P-sample and designed to support the estimation of erroneous enumerations. The data processing included a computerized search of census records to identify census enumerations for the P-sample and E-sample individuals (Cantwell et al. 2009). In addition, a computer-assisted clerical operation searched for enumerations for the P-sample individuals in the local area as well as duplicates of E-sample enumerations. When there was ambiguity, fieldwork collected additional information to resolve the status. Each P-sample and E-sample record that CCM processed was assigned a residence code indicating one of the following: (1) the person was a resident of the sample block cluster on Census Day, (2) was not a resident on Census Day, or (3) had unresolved Census Day residence. Figure 1 displays an overview of the CCM data collection and processing.

The P-sample interviews occurred in August and September 2010 independently from the 2010 Census. These interviews collected data that enabled constructing the Census Day (April 1) roster for the address by asking when current residents moved to the address

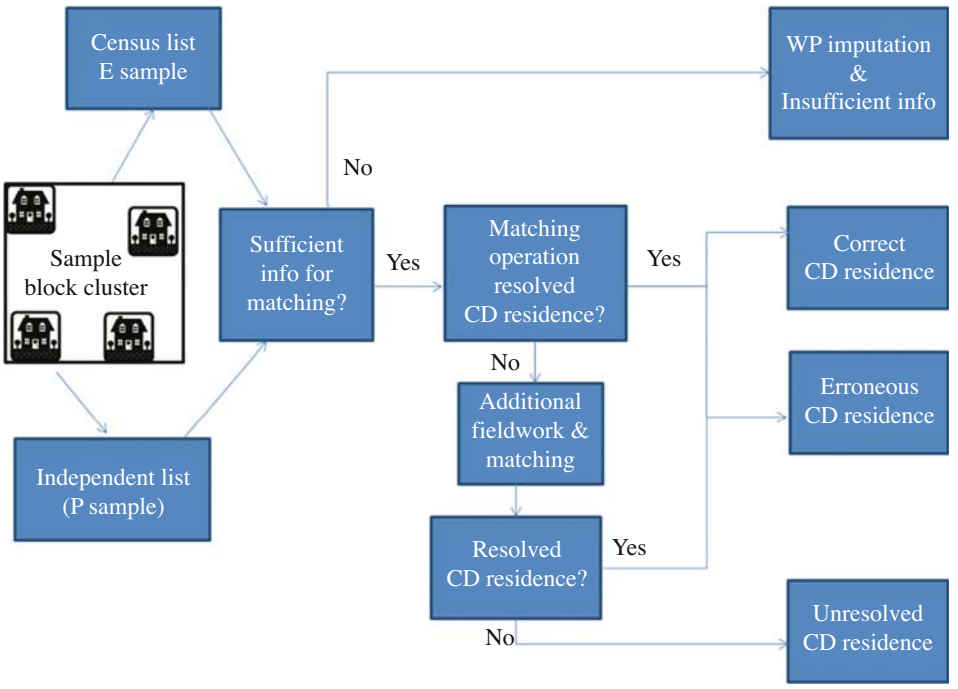

Fig. 1. Overview of CCM data collection and processing that produces codes indicating residence status on Census Day (CD). Note: WP imputation indicates whole person imputation, which is discussed in Subsection 2.4.

and about any Census Day residents who had moved from the address. The Census Bureau used a combination of electronic and clerical operations to match the P-sample people to the 2010 Census enumerations and conducted follow-up interviews in February 2011 to collect additional data when a person's Census Day residence could not be resolved. The CCM operation determined whether the census enumerations and P-sample persons were residents of their sample block cluster or the blocks surrounding the block cluster on Census Day by assigning the statuses of resident, nonresident, and unresolved. The CCM built this tolerance to avoid including minor geocoding error or mail delivery mistakes in the coverage error estimates, which would increase the variability of the estimates.

Since the P-sample is available only for the block clusters in the CCM sample, the comparison has to be restricted to the CCM block clusters. Although the 2010 CCM estimation does not require assuming that the P-sample interview is the 'truth,' the P-sample interviews are believed to be of higher quality because the interviewers have more training and experience since they were chosen from the pool of the best NRFU interviewers. In addition, the CCM interviewers were supported with a Computer Assisted Personal Interviewing (CAPI) instrument and supplied with additional residence probes.

The NRFU enumerations in the E-sample have residence status codes assigned during the CCM processing, but the administrative records in the NRFU housing units do not. We link the administrative records to the E- and P-sample records to retrieve CCM residence status codes. When a person's administrative record links to an enumeration in housing unit enumerated by a proxy response at the same address, the CCM residence code for the proxy response will indicate whether the person's enumeration at the address was correct. For example, if the person was enumerated at two addresses and the address not in the sample block was the correct Census Day residence, the enumeration in the sample block cluster was coded erroneous. This would mean the location of the person's administrative record was also in error. However, when a proxy response for a person and the administrative record file disagree, the CCM results provide information about whether the person should have been enumerated at the address and whether one of the sources is better for the person. Requiring the same address for a person's administrative record and the linking NRFU enumeration to retrieve a CCM residence code lends credibility to the assumption that the person lived at or is associated with the address. An administrative record will be inserted in the census at its address if the Census Bureau decides to use administrative records as enumerations. Requiring the same address from both sources means the correct enumeration rate reflects the accuracy of the use of administrative records at the addresses where they will be inserted in the census.

2.4. Matching Administrative Records to Combined CCM

The comparison of the 2010 Census NRFU housing units with proxy responses and the administrative records data for the housing units in the CCM block clusters requires linking the administrative records to the combined CCM to retrieve residence codes assigned during the CCM processing. The linking between the administrative records data and the combined CCM requires that both sources include Protected Identification Keys (PIKs). These PIKs are essentially encrypted Social Security Numbers or Individual Tax Identification Numbers, which are included when we use the term Social Security

Numbers. Administrative records data comes with Social Security Numbers that the Census Bureau staff converts to PIKs after a validation of their accuracy through matching to Social Security Administration files, a procedure called the Person Identification Validation System (PVS) (Wagner and Layne 2014). When a data file with records for persons does not come with Social Security Numbers, the Census Bureau uses its system to look up Social Security Numbers in Social Security Administration files and encrypt them by assigning PIKs. For this work, census and P-sample person records were assigned a PIK in a cascading search through the four search modules discussed in Wagner and Layne (2014): geographic search, name search, date of birth search, and household composition search. Each module has its own set of user defined blocking passes and parameter score thresholds. Layne et al. (2014) examine the error in PIK assignment by the PVS system associated with each of those search modules. It should be noted that this research assumes all PIKs are assigned with equal accuracy. PIKs have been assigned to the 2010 Census so the NRFU enumerations in the housing units with proxy responses have PIKs. PIKs also have been assigned to all the names collected in the P-sample regardless of the ultimate classification of nonmover, in-mover, out-mover, or never a resident of the sample block. Figure 2 illustrates the process of assigning PIKs and linking the files.

Sometimes the PVS fails to assign a PIK to a record. For example, 90.3% of the 2010 Census enumerations received a PIK from the PVS, but only 97% of the enumerations had enough information for an attempt to assign a PIK (Wagner and Layne 2014). Evaluation studies have shown that missing date of birth in a record is highly correlated with the PVS not assigning a PIK. In addition, an incomplete or fake name in a record is highly correlated with a PIK not being assigned (Wagner and Layne 2014; Mulrow et al. 2011). Nevertheless, it is possible to assign a PIK for someone missing sex or age particularly if other blocking and matching variables exist by which a high quality match can be made. However, a missing matching variable may result in a lower match score. Mulrow et al. (2011) found socioeconomic differences between the records that received PIKs and those that did not in a study using American Community Survey data. For example, the percentage assigned a PIK tended to be higher among those over 35 years of age than those younger. In addition, a higher percentage of those with a college degree received a PIK than those with a high school degree but not a college degree.

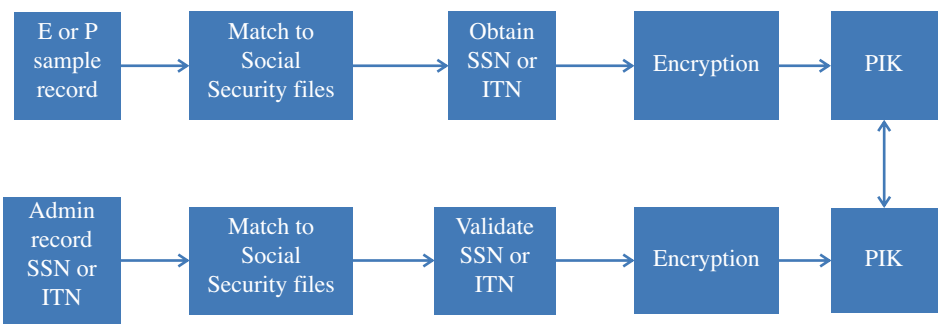

Fig. 2. The PVS assigns a Protected Identification Key (PIK) based on the person's Social Security Number (SSN) or Individual Tax Identification Number (ITN) for matching between the CCM E- and P-sample records and administrative records.

Having the CCM results available to compare proxy responses and administrative records is important because the estimated correct enumeration rate for the 2010 Census was 70.1% for persons enumerated by proxy respondents with 23.1% having all characteristics imputed, 5.6% being duplicates, and 1.1% being erroneous for other reasons. In contrast, 93.4% of the persons enumerated by a household member in NRFU were correct with 1.6% having all characteristics imputed, 4.2% being duplicates, and 0.8% being erroneous for other reasons (Mule 2012, Keller and Fox 2012). Even though enumerations that had all characteristics imputed, called *whole person imputations*, were not processed in the CCM E-sample due to lack of information to identify a person uniquely, the corresponding housing unit was included in the CCM P-sample and usually has information about the residents that can be used to evaluate any administrative records associated with the address. The P-sample also may have residency information for enumerations that are data-defined (i.e., processed in the E-sample) but have insufficient information to be processed in the CCM. The CCM requirement for sufficient information is a name and at least two characteristics because the CCM operations matched the enumerations to the names on the P-sample interview rosters.

When a person is enumerated by a proxy response and is in the administrative records file at the same address, the CCM residence code for the proxy response indicates whether the person's enumeration at the address was correct. If a person appears in the administrative records file but does not link to a combined CCM record at the same address, we can search the PIKs assigned to 2010 Census enumerations to learn if the person was enumerated elsewhere, but are not able to assess the accuracy for enumerations outside the CCM sample block clusters. If the person has an enumeration elsewhere that could not be assigned a PIK, we are not able to detect it using PIK matching.

Other types of electronic matching algorithms that do not rely on the assignment of PIKs, such as the household-based matching used by CCM, were not attempted. Household-based matching may or may not identify additional links between administrative records and the combined CCM. Regardless, our results must be viewed as conditional on the use of PIK matching.

Linking the administrative records to the CCM records enables identifying administrative records that are at the correct Census Day residence and those that are at an erroneous Census Day residence. Then a comparison of the percentages of administrative records and NRFU proxy responses in the CCM sample at the correct Census Day residence provides a measure to answer the research questions in Subsection 2.1.

2.5. *Underlying Assumptions*

This study approach has five major underlying assumptions:

- The results for proxy interviews in NRFU in the 2010 Census are applicable to the proxy interviews that would occur in the 2020 Census. The implementation of self-response and NRFU in the 2020 Census will be different from what occurred in the 2010 Census, and in particular, the procedures for taking proxy interviews in NRFU will differ.
- The 2010 CCM was able to determine whether the people on the rosters in NRFU proxy interviews were enumerated at the correct location, meaning their usual residence.

- The electronic matching algorithm used in this study (described in Subsection 2.4) was able to link a person's administrative record to the same person's record in the CCM P and E-samples.
- The availability of records from the administrative sources used in this study reflects the future availability from these sources.
- When a person has the same address in administrative records and NRFU, the person lives at or is associated with the address.

2.6. *Data*

For this study, we are going to focus on housing units in the CCM sample block clusters that were on the NRFU list in the E-sample and on the independent list of housing units created for the P-sample, and call this group the *combined CCM*. We need both E-sample and P-sample records because some or all the records for an occupied housing unit on the census list may be whole person imputations, but the P-sample interviewers were able to obtain data for the residents. In addition, the P-sample may have information regarding persons in administrative records not listed on the census form. We use the combined CCM to look up residence status codes for the administrative records. We do not form estimates using the combined CCM.

The administrative records file is the merger of the two files unduplicated within housing units: (1) the IRS 1040 forms filed in all months of 2010, (2) the Medicare records for all months of 2010. One reason the files were not unduplicated across housing units is that when duplicate records appear, there is no way to determine which is at the person's usual residence on Census Day. As stated earlier, the files from these two sources contain data for whole households. In addition, the 2014 Census Test operations used only these two sources.

The combined CCM contains 27,724 housing units that were proxy responses in NRFU with 10,416 occupied in NRFU, 15,012 vacant and 2,296 deleted because they did not have living quarters. Table 1 shows that of the 10,416 occupied housing units, 5,310 also have administrative records, the implication being that 5,106 have no records in the administrative records files we are using. Therefore, enumeration of these 5,106 housing units with proxy respondents using the combination of IRS 1040 and Medicare files is not an option unless other administrative sources with records for the housing units are found. However, one must keep in mind that the CCM oversamples hard-to-count areas. For a fit-for-use check, the percentage of the 23.6 million occupied housing units in NRFU that

Table 1. 2010 Census NRFU housing units in the combined CCM by administrative records (AR) status and type of NRFU respondent (unweighted).

	Proxy		HH Member	
AR status of housing units	HUs	%	HUs	%
Person records on AR list	5,310	51.0	16,876	61.3
No person records in AR list	5,106	49.0	10,647	38.7
Total	10,416	100.0	27,523	100.0

Note: Administrative records include IRS 1040 forms and Medicare records for all of 2010.

Table 2. Number of individual records found in administrative records (AR) files and number of individual records found on the combined CCM list in housing units in the combined CCM and occupied in the census by type of NRFU respondent.

Respondent type	AR	NRFU
Proxy	12,880	11,766
Household member	50,876	51,485
Total	63,756	63,251

have records in the combination of IRS 1040 and Medicare files is 56%. Therefore, the combined CCM percentages are reasonably comparable with proxy housing units being a little lower than the overall average at 51% and the housing units with household member respondents being a little higher at 61.3%.

For the NRFU housing units in Table 1 that have administrative records, Table 2 shows the distribution of the number of NRFU person records enumerated by proxy and household member respondents and the corresponding number of administrative records for the same housing units. In each of the two sources, the size of population in the proxy housing units is about 25% of the size of population in the housing units enumerated by household members. The administrative records file has more people in housing units enumerated by proxy than NRFU but fewer people in the housing units enumerated by household members. To see what would happen if all of these NRFU housing units were enumerated using administrative records, we combine the administrative records for NRFU housing units enumerated by both types of respondents and observe that the administrative records file has 505 records more than NRFU, about a 0.8% difference. Late in the analysis, we discovered that 88 of the administrative records persons in the proxy housing units and 237 in the housing units enumerated by a household member had died in 2009. These remain in the analysis but we address this issue for administrative records file construction in the recommendations in Section 4.

The 5,310 housing units with administrative records had 11,766 NRFU enumerations of persons with 9,258 of those having at least two characteristics, which is considered enough information to be an enumeration and is called *data-defined*. One of these characteristics could be a name. The remaining 2,508 were whole person imputations. Therefore, the imputation rate in these housing units is 21.3%, which is lower than the 23.1% for imputations among NRFU proxy enumerations nationally.

For completeness, we note that our analysis does not include 1,048 housing units with proxy respondents in the E-sample that are not also on the P-sample list, making them ineligible for the combined CCM list. The number of these housing units containing administrative records is 231 resulting in 460 administrative records for persons not being evaluated. In addition, the study does not include the 6,154 housing units on the P-sample list that were not on the E-sample list.

2.7. Evaluation Criteria

The evaluation of the quality of enumerations from the proxy responses and records in the administrative records file in the same housing units includes the rate of correct enumerations. The assessment also includes comparing the count of persons in each

source. Comparable calculations are made for enumerations and administrative records in housing units with household member responses.

- The total number of people enumerated at the sample addresses in each source.
- The total number of people correctly enumerated at the sample addresses in each source.
- HUs classified by (1) all administrative records are at the correct Census Day residence, (2) at least one administrative record is erroneous (not at the Census Day address) or its Census Day residence is unresolved, and (3) at least one Census Day resident does not have an administrative record at the address.

3. Results

Although the focus of our analyses is the NRFU housing units enumerated by proxy respondents, we are going to present results for NRFU housing units enumerated by household members for comparison. First, Subsection 3.1 considers the quality of the records for persons using the results of the CCM to determine whether the address on the record is in the correct location. Analyzing the quality of individual records provides insight when viewing the quality of the records for complete households, which is the focus of Subsection 3.2. In addition, analyses of individual records provide information about several potential uses of administrative records, such as for enumeration and for use in developing imputation models.

3.1. Quality of Individual Person Records

Even though Table 2 shows the number of records in administrative records and NRFU generally agree, this alone is not enough to evaluate the quality of the individual records in the two systems, which is the topic of our first research question. We need to know whether a person's record is at the correct location of the person's Census Day residence and whether the characteristics of the person and the size and composition of the households are correct.

Two things have to happen to evaluate an administrative record for a person: (1) the person's administrative records PIK has to link to the PIK for a record in the combined CCM and (2) the combined CCM record has to have a resolved residence status.

Table 3 shows the weighted distribution of combined CCM residence status for enumerations and administrative records in NRFU housing units in the combined CCM by NRFU respondent type while Table A1 in the Appendix shows the same results unweighted. The first thing to notice is that the unweighted and weighted distributions of CCM residence status are very similar for each NRFU respondent type. The weighted and unweighted distributions for the administrative records in housing units by NRFU respondent type also are similar. The weights are the CCM E-sample block cluster weights not adjusted for CCM nonresponse. Since the CCM sample design was able to keep the block cluster weights within a tight range, the similarity of the unweighted and weighted distributions is reasonable. We use the weighted results in our discussion.

To compare the distributions of the residence statuses from different types of respondents or different sources, we perform a chi-square test using the Rao-Scott adjustment (Lohr 1999) to account for the sampling design. For the design effect of the

Table 3. Weighted distributions of combined CCM residence status for enumerations and administrative records (AR) in NRFU housing units in the combined CCM by NRFU respondent type (shown in thousands).

	Proxy respondent			
	NRFU		AR	
Census Day residence status	count	%	count	%
Correct residence	5,235.2	56.6	5,017	49.1
Erroneous residence	380.9	4.1	418	4.1
Unresolved residence	1,462.4	15.8	379	3.7
NRFU not processed by CCM				
Insufficient info	258.3	2.8	-	-
Whole person imputation	1,920.6	20.7	-	-
AR PIK not in census at same address			4,397	43.1
Total	9,257.4	100.0	10,212	100.0
	Household member respondent			
	NRFU		AR	
Census Day residence status	count	%	count	%
Correct residence	36,720.2	88.0	29,971	72.5
Erroneous residence	1,058.9	2.5	1,054	2.5
Unresolved residence	2,308.2	5.5	1,283	3.1
NRFU not processed by CCM				
Insufficient info	1,070.9	2.6	-	-
Whole person imputation	583.0	1.4	-	-
AR PIK not in census at same address			9,038	21.9
Total	41,741.2	100.0	41,346	100.0

CCM sample, we examined Table 8 in Olson and Griffin (2012) that contains the means of several ranges of the observed correct enumeration rate, the number of observations in each range, and the standard error of the mean. The design effects varied between 2.5 and 3.5 across the categories. We use a design effect of three for the Rao-Scott adjustment to the chi-square statistics. For the chi-square tests, we use four cells: correct residence, erroneous residence, unresolved residence, and unable to process. For NRFU, we define the unable-to-process cell by collapsing insufficient information for CCM and whole person imputations. For administrative records, we collapse the records found at another census address and those not linked to a combined CCM record.

For the NRFU proxy enumerations, Table 3 shows that CCM found that 56.6% were at the correct residence, and 4.1% were at an erroneous residence. CCM attempted but could not determine Census Day residence for 15.8% of the NRFU proxy enumerations. CCM did not attempt to process the 2.8% that had insufficient information or the 20.7% that were whole person imputations.

For the NRFU enumerations by household members in Table 3, we see that 88.0% are at the correct residence, 2.5% are at an erroneous residence, and 5.5% had an unresolved residence status. However, 2.6% had insufficient information for CCM to process and 1.4% of the proxy enumerations were whole person imputations, which CCM did not process.

Turning to the residence status of the administrative records in NRFU housing units in Table 3, for proxy respondents, links to combined CCM records showed that 49.1% were

at the correct residence, 4.1% were at an erroneous residence, and 3.7% had an unresolved residence. The percentage that did not link at the same address and could not be evaluated is 43.1%. When we examine the administrative records in the housing units with household member respondents, we see that links to the combined CCM found that 72.5% were at the correct residence, 2.5% were at an erroneous residence, and the residence status of 3.1% could not be resolved. The percentage that did not link at the same address and could not be evaluated is 21.9%.

For some insight about the administrative records that did not link, the unweighted data in Table A1 shows that 17.3% of the individual enumerations by a proxy respondent and 10.5% of the individual enumerations by a household member respondent did not link to a combined CCM record at the same address but linked to enumerations elsewhere in the census. In addition, 26.8% of the individual enumerations by a proxy respondent and 12.9% of the enumerations by a household member respondent did not link to a combined CCM record at the same address or elsewhere in the census. For the administrative records found elsewhere in the census, using the administrative records for enumeration would create duplicate enumerations. We do not have information to determine which address was the correct location for their enumeration since the census address was not in the CCM. These people may have moved or may alternate between two residences, such as families with seasonal homes or children in shared custody. In these cases, they may have been enumerated in one location and list the other address as their residence in administrative records. As for the administrative records that did not link anywhere in the census, there are two possible explanations: (1) the person has a census enumeration but it has errors or not enough information for the linking procedure to make the connection; (2) the person was missed by the census.

Next, we compare the distributions of the residence statuses for the NRFU enumerations and the administrative records by respondent. For the housing units with proxy respondents, the chi-square test produced a p-value less than 0.001, which leads us to conclude that the distribution of the residence statuses for the NRFU enumerations with 56.6% correct and the administrative records with 49.1% correct are different. For the housing units with household member respondents, the p-value of the chi-square test is 0.028, which indicates the distributions of the residence codes are different. For both types of respondents, the percentage of NRFU enumerations at the correct residence is higher than observed for administrative records, and the percentage of administrative records that cannot be evaluated is higher than observed for NRFU enumerations.

Both NRFU and administrative records have a substantial percentage of records where this approach is unable to evaluate their residence status. The seemingly high percentage of records that do not link to a combined CCM record at their administrative records address but link to a census address elsewhere causes concern that these administrative records are not at the correct Census Day residence and more importantly, that inserting them as census enumerations would create duplicate enumerations. Since the CCM sample did not include the address where administrative records PIKs were found, the CCM did not evaluate the accuracy of the enumeration of the people at the address. Therefore, the accuracy of administrative records that linked to these enumerations also could not be evaluated.

Interestingly, the percentage of records with a CCM resolved residence status is higher for NRFU enumerations than administrative records in housing units with both

types of respondents. Keep in mind that all the administrative records have PIKs, but the Census Bureau procedure may or may not be able to assign PIKs to the census enumerations.

From another perspective, we compare the distributions of the residence status of the NRFU enumerations for the two types of respondents. A chi-square test comparing produced a p-value less than 0.001; therefore, we conclude that the distributions are different. We see that the percentage of proxy enumerations that are at the correct residence at 56.6% is lower than the percentage of household member enumerations at the correct residence at 88.0%. The most apparent difference is that the percentage of whole person imputations is much higher for the proxy enumerations at 20.7% than for the household member respondents at 1.4%. However, the housing units that are remaining after the attempts to get household member respondents fail get rolled over to the attempts to get proxies. So, almost all the whole person imputations are attributed to the proxies, although both the self-response phase and the NRFU household member response phase also fail to get a response.

Similarly, a chi-square test to compare the distributions of the administrative records for the two respondent types produces a p-value of 0.010, which indicates that the distributions are different. The percentage of administrative records that are at the correct residence is 49.1% in the housing units enumerated by proxy while the percentage correct is higher at 72.5% in the housing units enumerated by a household member. In addition, the percentage that did not link at the same address and could not be evaluated is higher for proxy respondents 43.1% than for household member respondents at 21.9%.

3.2. Characteristics Correlated with Quality

When we consider our second research question, we note that the assignment of PIKs to the combined CCM records proved crucial to evaluating the administrative records in housing units enumerated during NRFU. Therefore, the percentage of NRFU enumerations that received PIKs is an evaluation tool. Table 4 shows the distribution of the residence status of enumerations with PIKs and those without PIKs by NRFU respondent. Of the NRFU enumerations where the PVS attempted to assign PIKs, 73% (SE = 0.9%) of those in housing units enumerated by proxy received PIKs while 92% (SE = 0.2%) of those enumerated by a household member received PIKs. If the whole person imputations are included, the percentage is 58% (SE = 0.8%) for proxy respondents and 91% (SE = 0.2%) for household member respondents. When whole person imputations are included and when they are not, the tests of difference between the percentages of enumerations assigned PIKs for proxy and household member respondents produced p-values less than 0.001, so we conclude there is a difference in the enumerations from the two types of respondents.

In summary, a distinguishing feature that indicates the quality of NRFU enumerations appears to be whether they can be assigned a PIK. Those that receive PIKs tend to be in the correct location at high rate. Table 5 shows the correct enumeration rate for several criteria for the denominator for enumerations with and without PIKs by type of NRFU respondent. We do not conduct statistical testing but use the data in Table 5 to illustrate the effect of the choice of the denominator of the correct enumeration rate.

Table 4. Weighted distributions of combined CCM residence status for enumerations in NRFU housing units by NRFU respondent type and PIK status (shown in thousands).

	Proxy			Household member		
Census Day residence status	with PIK	without PIK	Total	with PIK	without PIK	Total
PIK attempted						
Correct residence	3,625.8	1,609.4	5,235.2	34,322.1	2,398.2	36,720.2
Erroneous residence	266.4	114.5	380.9	844.0	214.9	1,058.9
Unresolved residence	337.5	173.2	510.7	1,713.5	594.7	2,308.2
Insufficient info for CCM	1,124.9	85.1	1,210.0	990.8	80.1	1,070.9
Subtotal	5,354.6	1,982.1	7,336.7	37,870.3	3,287.9	41,158.3
	73%	27%	100%	92%	8%	100%
PIK not attempted						
Whole person imputation		1,920.6	1,920.6		583.0	583.0
Total	5,354.6	3,902.8	9,257.4	37,870.3	3,870.9	41,741.2
	58%	42%	100%	91%	9%	100%

Table 5. Weighted correct enumeration (CE) rate for enumerations in occupied housing units in the combined CCM with several criteria for the enumerations included in the denominator by type of NRFU respondent. (shown in thousands).

Status of enumerations in denominator	Proxy respondent			HH member respondent		
	Total	CE	% CE	Total	CE	% CE
With PIK						
CCM resolved status	3,892	3,626	93	35,166	34,322	98
Data-defined	5,355	3,626	68	37,870	34,322	91
Without PIK						
CCM resolved status	1,724	1,609	93	2,613	2,398	92
Data-defined	1,982	1,609	81	3,288	2,398	73
Data-defined and imputed	3,903	1,609	41	3,871	2,398	62

When the denominator includes only the enumerations where CCM could resolve the residence status, namely those that are correct and erroneous, the percentage correct is not dramatically different from the percentages for the household member respondents without PIKs and both categories for proxy respondents, which range from 92% to 98%. Additionally, Table 3 shows that the percentage of administrative records with a resolved residence status in proxy housing units that are correct is in the same range at 92% (5,017/(5,017+418)).

For the data-defined enumerations with PIKs, 68% from proxy respondents and 91% from household member respondents are in the correct location. However, the correct enumeration rate among enumerations that are data-defined but not assigned a PIK is 81% for proxy respondents and 73% for household member respondents. When the denominator for those without PIKs includes whole person imputations, the correct enumeration rate for proxy respondents is 41%. For household member respondents, rate becomes 62% with the inclusion of the imputations. Keep in mind that whole person imputations are a much smaller percentage of the enumerations by household members than for proxy respondents.

3.3. Quality of Records for Entire Households

As stated in the third research question, our ultimate interest is the quality of administrative records on a household basis because that is most likely the way they will be used for enumeration. Our analysis examines two measures. One is the percentage of housing units where the population counts from NRFU and administrative records are equal. The other is the percentage of NRFU housing units where the combined CCM determines the administrative records roster is perfect. These are descriptive analyses with unweighted data.

Table 6 shows that the percentage housing units where the NRFU and administrative records population counts are the same is 51% for both proxy and household member respondents. However, the administrative records population count being equal to the NRFU population count does not mean that the administrative records roster for the housing unit has the correct Census Day residents. CCM provides a means to determine the accuracy of the administrative records roster.

Table 6. Unweighted comparison of housing unit population counts from NRFU and administrative records (AR) by respondent type.

	Proxy		Household member	
Housing unit population counts	Number of housing units	%	Number of housing units	%
Same AR and census	2,685	51	8,633	51
Different AR and census	2,625	49	8,243	49
Total	5,310	100	16,876	100

Therefore, we examine the accuracy of the administrative records on a household basis for the 5,310 housing units with proxy respondents and 16,876 housing units with household member respondents that have administrative records. Table 7 shows the percentage of housing units in the following categories as determined by the combined CCM:

- Administrative Records Perfect – All administrative records persons in the housing units are Census Day residents at the address and no Census Day residents are omitted from the administrative records roster.
- Administrative Records Erroneous Enumerations and Unresolved Enumerations (E&Us) – At least one administrative record in the housing unit either linked to a combined CCM record coded as not being a Census Day resident at the address or did not link to a combined CCM record with a resolved residence status.
- Administrative Records Omissions – There is at least one person that the combined CCM found to be a Census Day resident at the address, but the person(s) is (are) not on administrative records roster for the address.

When the administrative records in the 5,310 proxy housing units are considered on a household basis instead of a individual basis, 1,722 (32.4%) are perfect in that the combined CCM indicated every record as being at the person's Census Day residence and no persons were omitted. We also find that administrative records for 408 (7.7%) of the housing units omit at least one person that the combined CCM found to be a Census Day resident at the address. The remaining 3,180 (59.9%) have at least one record that the combined CCM found not to be a resident at the address on Census Day, or the person's Census Day residence was not determined because the administrative records did not link to a combined CCM record with a resolved residence status.

Table 7. Status of administrative records (AR) in NRFU housing units in the combined CCM by NRFU respondent type (unweighted).

	Proxy		Household member	
Housing unit status	Number of housing units	%	Number of housing units	%
AR Perfect	1,722	32.4	7,256	43.0
AR E&U	3,180	59.9	6,846	40.6
AR Omissions	408	7.7	2,774	16.4
Total	5,310	100.0	16,876	100.0

Surprisingly, the percentage of housing units with household member respondents who omitted at least one Census Day resident from the administrative records roster was 16.4%. In addition, 43.0% of the administrative records rosters for housing units enumerated by household members are perfect. The percentage of housing units with an administrative record for at least one person who was not a Census Day resident or had an unresolved Census Day residence was 40.6%.

4. Summary

Our investigation discovered that determining whether proxy responses are more or less accurate than administrative records is not as straightforward as it sounds. The percentage of enumerations in housing units with proxy respondents in the correct location units was higher than the percentage for administrative records in the same housing units even though the administrative records sources were all IRS 1040 and Medicare records from 2010. However, the percentage of records that could not be evaluated was higher for the administrative records than for the proxy respondents. The high unresolved rate among administrative records was due to the failure to link the administrative records to a combined CCM record at the same address. The reasons that an administrative record did not link include the individual being enumerated at another address, having a census enumeration or P-sample roster entry that could not be assigned a PIK, or being missed by the census. This research prompted a change from the initial plan that used all administrative records for NRFU enumeration to the search for methods to identify the best administrative records for enumeration. The current methodological approach focuses on the development of predictive models to identify administrative records with a high probability of being accurate (Morris et al. 2016).

In addition, the findings of our study have implications for the census in the areas of administrative records sources used in census enumeration, the risk of duplication, and characteristics of high quality proxy enumerations. We recommend finding additional high-quality administrative records sources to increase the potential for using administrative records to enumerate housing units that cannot be enumerated well by proxy, such as the Supplemental Nutrition Assistance Program files from the states. We found that enumeration with administrative records was not an option for approximately half of the housing units in the CCM E-sample classified as occupied in the census using proxy respondents when the administrative records sources were IRS 1040 and Medicare files for all of 2010. If additional high quality administrative records sources cannot be found, these housing units without administrative records will need to be contacted by NRFU enumerators or imputed. However, the implication of increasing the number of administrative records sources is that it elevates the importance of identifying duplicate records across housing units and developing rules for which address to keep as the person's Census Day address. Algorithms for identifying duplicate records face challenges when sources do not have the same name, age, and/or address for a person. Examples include when one source has a person's nickname while the other source has the given name, and the old versus new address when a person moves. Adding sources of administrative records has the potential to increase the variation in the key variables used in linking the records, thereby increasing the errors in identifying duplicates.

One important finding is that not all proxy responses are bad as demonstrated by the result that over half of proxy enumerations are in the correct location (56.6%). By almost any standard, proxy enumerations that can be assigned PIKs tend to be in the correct location. Therefore, one indicator for a high quality NRFU enumeration appears to be whether it has enough information for the Census Bureau's Personal Validation System algorithm to assign a PIK. The implication is that the design of NRFU operations would profit by including strategies to obtain high-quality proxy responses. Such strategies would include designing the training of interviewers to emphasize the importance of obtaining the name and age of the residents from proxy respondents since these are important for assigning PIKs. Additional advantages may come from developing contact tactics that incorporate the times when knowledgeable proxy respondents are likely to be accessible, namely at home for neighbors or on the premises for multi-unit building managers.

However, the amount of information collected for an individual does not always assure that the PVS will be able to assign a PIK. Some data-defined census enumerations that meet the CCM criteria of sufficient information, which is a name and two characteristics, could not be assigned PIKs but were found by CCM to be enumerated at the correct location. The addresses where these people were enumerated may not have been associated with them in administrative records.

Since administrative records enumeration would occur on a housing unit basis, a comparison of NRFU proxy responses and administrative records for whole households on population count and accuracy of location also is important. The combined CCM found that an unweighted 32% of the administrative records for proxy housing units were enumerated perfectly. That means that all the administrative records persons in the housing unit were Census Day residents and no Census Day residents were omitted from the administrative records roster. The enumerations with unresolved residence status were not considered to be at the correct location. Some likely are, but without enough information to make a determination. When focusing only on population count, the percentage of housing units have an administrative records count that agrees with the census count is an unweighted 51% among housing units with proxy respondents and among housing units with household member respondents.

The results also indicate that duplication may be a problem when using administrative records to enumerate whole HHs. Census operations may need to search census enumerations, particularly self-responses, to be sure that an administrative records enumeration does not create a duplicate. If a search finds another enumeration for a person, the administrative record is not necessarily the one in the wrong location. Self-responses may be in error due to postal delivery errors or misunderstandings about the correct location for enumeration when a person has more than one residence. One approach to identifying which of two enumerations to keep in the census is to consult multiple administrative records sources and make the decision based on the recency and frequency of the appearances of the person at the addresses. The addition of questions regarding other residences to the census questionnaire may aid in avoiding duplicates.

Further research is needed to identify additional characteristics that indicate how the quality of the proxy responses may vary. Additional investigations could examine the demographic, geographic, and socioeconomic characteristics of the housing units where the combined CCM found their individual administrative records to be perfect, that is, the

exact household members were correctly enumerated versus those housing units with administrative records that had errors or could not be evaluated. Additional research could examine relationships between operational characteristics, such as the number of prior contact attempts and correct proxy responses to identify characteristics of housing units with complete correct administrative records among NRFU proxy responses.

The results of our study apply to identifying a person's usual residence and characteristics for census-taking and therefore probably have only limited implications for surveys since the focus of surveys usually is to collect behavior or socioeconomic information. Survey researchers do need to be aware that when linking a survey from a sub-national area to append additional administrative records data to individual records, the respondents may not be in administrative records at the survey address. As for administrative records, our study indicates that while administrative records contain a large amount of information, determining whether that data is truly adequate for the purpose at hand is not always easy.

Appendix

Table A1. Unweighted distributions of combined CCM residence status for enumerations and administrative records (AR) in NRFU housing units in the combined CCM by NRFU respondent type.

	Proxy respondent			
	NRFU		AR	
Census Day residence status	count	%	count	%
Correct residence	6,637	56.4	6,191	48.1
Erroneous residence	481	4.1	519	4.0
Unresolved residence	1,850	15.7	493	3.8
NRFU not processed by CCM				
Insufficient info	290	2.5	-	-
Whole person imputation	2,508	21.3	-	-
AR PIK not in census at same address				
Found at another census address	-	-	2,230	17.3
Not linked to census records	-	-	3,447	26.8
	11,766	100.0	12,880	100.0
	Household member respondent			
	NRFU		AR	
Census Day residence status	count	%	count	%
Correct residence	45,018	87.4	36,084	70.9
Erroneous residence	1,392	2.7	1,258	2.5
Unresolved residence	3,042	5.9	1,645	3.2
NRFU not processed by CCM				
Insufficient info	1,285	2.5	-	-
Whole person imputation	748	1.5	-	-
AR PIK not in census at same address				
Found at another census address	-	-	5,318	10.5
Not linked to census records	-	-	6,564	12.9
	51,485	100.0	50,869	100.0

5. References

Cantwell, P.J., M. Ramos, and D. Kostanich. 2009. "Measuring Coverage in the 2010 U.S. Census." In *JSM Proceedings*, Social Statistics Section, American Statistical Association, Washington, DC, August 1–6, 2009. Alexandria, VA: American Statistical Association. 43–54. Available at: https://ww2.amstat.org/sections/srms/proceedings/y2009/Files/302739.pdf (accessed January 2017).

Chesnut, J. 2005. "Item Nonresponse Error for the 100 Percent Data Items on the Census 2000 Long Form Questionnaire." In *JSM Proceedings*, Section on Survey Research Methods, American Statistical Association, Minneapolis, MN, August 7–11, 2005. Alexandria, VA: American Statistical Association. 2857–2864. Available at: http://ww2.amstat.org/sections/srms/Proceedings/y2005/Files/JSM2005-000341.pdf (accessed January 2017).

Keller, A. and T. Fox. 2012. "2010 Census Coverage Measurement Estimation Report: Components of Census Coverage for the Household Population in the United States." DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-G-04. Washington, DC: U.S. Census Bureau. Available at: http://www.census.gov/coverage_measurement/pdfs/g04.pdf (accessed January 2017).

King, T., S. Cook, and J. Hunter Childs. 2012. "Interviewing Proxy Versus Self-Reporting Respondents to Obtain Information Regarding Living Situations." In *JSM Proceedings*, Survey Research Methods Section, American Statistical Association, San Diego, CA, July 28–August 2, 2012. Alexandria, VA: American Statistical Association. 5667–5677. Available at: https://ww2.amstat.org/sections/srms/proceedings/y2012/files/400243_500698.pdf (accessed January 2017).

Layne, M., D. Wagner, and C. Rothhaas. 2014. "Estimating Record Linkage False Match Rate for the Person Identification Validation System." CARRA Working Paper Series. Working Paper #2014-02. Washington, DC: Census Bureau. Available at: https://www.census.gov/library/working-papers/2014/adrm/carra-wp-2014-02.html (accessed March 2017).

Lohr, S. 1999. *Sampling: Design and Analysis*. Cengage Learning. Boston, MA.

Martin, E. 1999. "Who Knows Who Lives Here? Within-household Disagreements as a Source of Survey Coverage Error." *Public Opinion Quarterly* 63: 220–236. Doi: http://dx.doi.org/10.1086/297712.

Morris, D., A. Keller, and B. Clark. 2016. "An Approach for Using Administrative Records to Reduce Contacts in the 2020 Census." *Statistical Journal of the IAOS* 32: 177–188. Doi: http://dx.doi.org/10.3233/SJI-161002.

Mule, T. 2012. "Census Coverage Measurement Estimation Report: Summary of Estimates of Coverage for Persons in the United States." DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-G-01. Washington, DC: U.S. Census Bureau. Available at: http://www.census.gov/coverage_measurement/pdfs/g01.pdf (accessed January 2017).

Mulrow, E., A. Mushta, S. Pramanik, and A. Fontes. 2011. Assessment of the U.S. Census Bureau's Person Identification Validation System. Report for the U.S. Census Bureau. Chicago, IL: NORC. Available at: http://www.norc.org/PDFs/May%202011%20Personal%20Validation%20and%20Entity%20Resolution%20Conference/PVS%20Assessment%20Report%20FINAL%20JULY%202011.pdf (accessed January 2017).

Mulry, M.H. and B.D. Spencer. 2012. "A Framework for Cost Models Relating Cost and Data Quality." Presentation at the 2012 International Total Survey Error Workshop. Sanpoort, The Netherlands, September 2–4, 2012. Research Triangle Park, NC: National Institute of Statistical Science. Available at: http://www.niss.org/sites/default/files/Mulry_september2012.pdf (accessed January 2017).

Olson, D. and R. Griffin. 2012. "2010 Census Coverage Measurement Estimation Report: Aspects of Modeling." DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-G-10. U.S. Washington, DC: Census Bureau. Available at: http://www.census.gov/coverage_measurement/pdfs/g10.pdf (accessed January 2017).

U.S. Census Bureau. 2015. *Planning Database*. Washington, DC: Census Bureau. Available at: http://www.census.gov/research/data/planning_database/ (accessed January 2017).

Wagner, D. and M. Layne. 2014. "The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications." CARRA Working Paper Series. Working Paper #2014-01. Washington, DC: Census Bureau. Available at: https://www.census.gov/library/working-papers/2014/adrm/carra-wp-2014-01.html (accessed March 2017).

Wolfgang, G., R. Byrne, and S. Spratt. 2003. *Analysis of Proxy Data in the Accuracy and Coverage Evaluation*, Census 2000 Evaluation O.5. Washington, DC: U.S. Census Bureau. https://www.census.gov/pred/www/rpts/O.5.PDF (accessed March 2017).

Received January 2016
Revised February 2017
Accepted March 2017