Are Proxy Responses Better Than Administrative Records?

Mary H. Mulry and Andrew Keller[1]
U.S. Census Bureau, Washington, DC 20233

Abstract

Currently the U.S. Census Bureau is 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. In previous censuses, 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, 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). 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, Nonresponse Followup, 2010 Census Coverage Measurement

1. Introduction

Currently the U.S. Census Bureau is 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 where the Census Bureau did not receive a self-response electronically or by mail. Regardless of the number of contact attempts the 2020 Census NRFU design permits, enumerators will confront the problem of not being able to contact the residents at some addresses. In previous censuses, the strategy at this point has been to find a knowledgeable person, such as a neighbor or apartment manager, who could provide the census information for the residents, called a proxy response. The Census Bureau's recent advances in merging federal and third-party databases to create households that can be used for census enumeration purposes raises the question: Are proxy responses for NRFU addresses more or less accurate than the administrative records available for the housing unit?

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. The comparison of the quality of the two sources uses the results of the 2010 Census Coverage Measurement (CCM). The goals of our study also include examining whether the quality of proxy responses for NRFU addresses vary by the number of contact attempts prior to the proxy response and/or by whether the administrative records available for the address are deemed high quality or low quality, with

[1] 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.

defining high quality as part of the research. The evaluation of the quality of the proxy responses and the administrative records files includes comparisons of the number of people enumerated, the number of people correctly enumerated, and the demographic distributions. 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.

The Census Bureau is conducting a series of tests to examine the implementation of adaptive strategies for conducting Nonresponse Followup (NRFU) of the housing units that do not self-respond in a census. The proposed strategies include using administrative records and a variable number of contact attempts with the goal of reducing costs and improving data quality.

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

Instead, the plan is to compare the quality of the 2010 Census NRFU HUs with proxy responses and the AR data for those HUs 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).

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 ARs in corresponding HUs. Another aspect is to use decision trees in developing statistical models to identify the characteristics of NRFU HUs 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 socio-economic 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

The focus of our research is to answer the following questions to produce information useful for the design of the strategy for contacting HUs during the 2020 Census NRFU:

- Are proxy responses for NRFU addresses more accurate than the administrative records available for the housing unit or are they less accurate?
- Does the quality of proxy responses for NRFU addresses vary by the number of contact attempts prior to the proxy response and/or by whether the administrative records available for the address are deemed high quality or low quality?

2.2 Population

The population under study is defined as the people whose Census Day residence is in a housing unit enumerated in the 2010 Census NRFU by a proxy respondent, and administrative records are available for the housing unit. 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. We consider the quality of two lists of the population using the criteria of whether the person is found on the list 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 (HH) member and the administrative records at these addresses.

In this study, the definitions of the populations enumerated by proxy and HH member respondents are operational and depend on the conduct of the 2010 Census operations. The HUs enumerated by HH member respondents failed to self-respond by mail. The HUs enumerated by proxy failed to self-respond by mail, and none of the HH 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 enumeration sample (E sample) and the population sample (P sample). The E sample and the P sample used the same sample of block clusters. The E sample contained the census enumerations in the block clusters and its design supported the estimation of erroneous enumerations. The P sample constructed its list of the population in the block clusters independently of the census and was designed to support the estimation of census omissions. 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.

The P sample interviews were conducted 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 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 determines whether the census enumerations and P-sample persons were residents of their sample block cluster on Census Day assigning the statuses of resident, nonresident, and unresolved. 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 given additional residence probes to ask.

The NRFU enumerations in the E sample have residence status codes assigned during the CCM processing, but the administrative records in the NRFU HUs 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 HU 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 therefore, whether one of the sources is better for the person.

2.3 Underlying assumptions

This study approach has four major underlying assumptions:

- One assumption is that 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.
- A second assumption is that 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 third assumption is that the electronic matching algorithm used in the study (described in Section 2.5) was able to link a person's administrative record to the same person's record in the combined CCM.
- The fourth assumption is that the availability of records from the administrative sources used in this study reflects the future availability from these sources.

2.4 Data

For this study, we are going to focus on HUs in the CCM sample block clusters that were on the NRFU list in the E sample and on the independent list of HUs 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 HU 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 ARs 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 unduplicated merger of the two files: (1) the IRS 1040 forms filed in all months of 2010, (2) the Medicare records for all months of 2010. The files from these two sources have appeared to be better than other sources in providing whole households. In addition, the 2014 Census Test operations used only these two sources.

The combined CCM contains 27,724 HUs that were proxy responses in NRFU with 10,416 occupied in NRFU, 15,012 vacant and 2,296 deleted because they did not having 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 AR files we are using. For comparison, the percentage of the 144,000 occupied HUs in NRFU that have records in the combination of IRS 1040 and Medicare files is 56%, which means the combined CCM percentages are reasonable with proxy HUs being a little lower at 51% and the HH member HUs being a little higher at 61.3%.

Table 1. 2010 Census NUFU HUs in the combined CCM by AR status and type of NRFU respondent (unweighted)

AR status of HUs	Proxy		HH Member	
	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: ARs include IRS 1040 forms and Medicare records for all of 2010.

For the NRFU HUs In Table 1 that have AR records, Table 2 shows the distribution of the number of NRFU person records enumerated by proxy and HH member respondents and the corresponding number of records for the same HUs. In each of the two sources, the size of population in the proxy HUs is about 25% of the size of population in the HH member HUs. The AR file has more people in HUs enumerated by proxy than NRFU but fewer people in the HUs enumerated by HH members. Combining all the NRFU HUs, the AR file has 505 records more than NRFU, about a 0.8% difference.

Table 2. Number of records found in AR files and number of record founds on the combined CCM list in HUs in the combined CCM and occupied in the CUF by type of NRFU respondent.

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

The 5,310 HUs with ARs had 11,766 NRFU enumerations of persons with 9,258 that had at least two characteristics, one of which could be a name, which was considered enough information to be an enumeration, called *data-defined*. The remaining 2,508 were whole person imputations. Therefore, the imputation rate in these HUs is 21.3%, which is lower than the national average of 23.1% for imputations among NRFU proxy enumerations.

For completeness, we note that our analysis does not include 1,048 HUs 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 HUs containing ARs is 231 resulting in 460 ARs for persons not being evaluated. In addition, the study does not include the 6,154 HUs on the P sample list that were not on the E sample list.

2.5 Matching ARs to combined CCM

The comparison of the 2010 Census NRFU HUs with proxy responses and the AR data for the HUs in the CCM block clusters requires linking the AR records to the combined CCM to retrieve residence codes assigned during the CCM processing. The linking between the AR data and the combined CCM requires that both sources have Protected Identification Keys (PIKs), which are essentially encrypted Social Security Numbers (SSNs) or Individual Tax Identification Numbers (ITNs, included when we use the abbreviation SSN. AR data comes with SSNs that the Census Bureau staff converts to PIKs after a validation of their accuracy through matching to Social Security Administration (SSA) 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 SSNs, the Census Bureau uses its system to look up SSNs in SSA files and encrypt

them by assigning PIKs. PIKs have been assigned to the 2010 Census so the NRFU enumerations in the HUs 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.

Having the CCM results available to compare the proxy responses and AR 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 HU was included in the CCM P-sample and will usually have information about the residents that can be used for evaluating an AR records associated with the address. The P sample also may have residency information for enumerations that were data-defined but had insufficient information to be processed in the CCM. The CCM requirement for sufficient information was 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 was enumerated by a proxy response and in the AR 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 AR 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 ARs and the combined CCM. Regardless, our results must be viewed as conditional on the use of PIK matching.

2.6 Evaluation criteria

The evaluation of the quality of enumerations from the proxy responses and records in the AR file in the same HUs 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 AR records in HUs with HH 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.
- Of the people where the two sources agree, the total number correctly enumerated and the total number erroneously enumerated.

3. Results

Although the focus of our analyses is the NRFU HUs enumerated by proxy respondents, we are going to present results for NRFU HUs enumerated by household (HH) members for comparison. First, Section 4.1 considers the quality of the records for persons under the criteria of whether the address on the record is in the correct location as determined by CCM. Analyzing the quality of individual records provides insight when viewing the quality of the records for complete

households, which is the focus of Section 4.2. 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 Analysis of individual person records

Even though Table 2 shows the number of records in ARs and NRFU generally agree, this alone is not enough to evaluate the quality of the individual records in the two systems. 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 AR for a person: (1) the AR PIK has to link to a record in the combined CCM, (2) the combined CCM record has to have a resolved residence status.

Table 3 shows the unweighted distribution of combined CCM residence status for enumerations and ARs in NRFU HUs in the combined CCM by NRFU respondent type while Table 3W shows the same results weighted. The first thing to notice is that the unweighted and weighted distributions of CCM residence status is very similar for each NRFU respondent type. The weighted and unweighted distributions for the ARs in HUs 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 with in 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 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 equal to 3 for the Rao-Scott adjustment to the chi-square statistics. In addition, 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, and for ARs, we collapse the records found at another census address and those not linked to a census record.

For the NRFU proxy enumerations, Table 3W 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 from HH members in Table 3W, 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.

A chi-square test comparing the distributions of the residence status of the NRFU enumerations for the two types of respondents produced a p-value less than 0.001, and therefore, we conclude that the distributions are different. We see that the percentage of proxy enumerations that are at the correct residence 56.6% is lower at than the percentage of HH member enumerations 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 HH member respondents at 1.4%. However, the HUs that are remaining after the attempts to get HH member respondents fail get rolled over to the attempts to get proxies so virtually all the whole person imputations get attributed to the proxies, although both the self-response phase and the NRFU HH member phase also failed to get a response.

Turning to the residence status of the ARs in NRFU HUs with proxy respondents in Table 3W, 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%. For some insight about the ARs that did not link, the unweighted data in Table 3 shows that 17.3% did not link to a combined CCM record at the same address but linked to an enumerations elsewhere in the census while 26.8% did not link to a combined CCM record at the same address or elsewhere in the census.

Table 3. Unweighted distributions of combined CCM residence status for enumerations and ARs in NRFU HUs in the combined CCM by NRFU respondent type

	Proxy respondent			
Census Day residence status	NRFU		AR	
	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%

	HH member respondent			
Census Day residence status	NRFU		AR	
	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%

Table 3W. Weighted distributions of combined CCM residence status for enumerations and ARs in NRFU HUs in the combined CCM by NRFU respondent type (**shown in 1,000's)**

	Proxy respondent			
Census Day residence status	NRFU		AR	
	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	2.8%	-	-
Whole person Imputation	1,903	20.7%	-	-
AR PIK not in census at same address			4,397	43.1%
Total	9,257	100.0%	10,212	100.0%

	HH member respondent			
Census Day residence status	NRFU		AR	
	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	258	2.6%	-	-
Whole person Imputation	1,903	1.4%	-	-
AR PIK not in census at same address			9,038	21.9%
Total	41,741	100.0%	41,346	100.0%

When we examine the ARs in the HUs with HH 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 29.1%. The unweighted data in Table 3 shows that 10.5% did not link to a combined CCM record at their AR addresses but were found at other addresses in the census while 12.9% did not link to the combined CCM at their AR addresses or at another address in the census

The chi-square test to compare the distributions of the ARs for the two respondent types produces a p-value of 0.010, which indicates that the distributions are different. The percentage of ARs that are at the correct residence is 49.1% in the HUs enumerated by proxy while the percentage correct is higher at 72.5% in the HUs enumerated by a HH member. There is not much difference in the percentages of the ARs that at an erroneous address or with an unresolved residence. However, the percentage that did not link at the same address and could not be evaluated is higher for proxy respondents 43.1% than for HH member respondents at 21.9%.

Next, we compare the distributions of the residence statuses for the NRFU enumerations and the ARs by respondent. For the HUs 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 and the ARs in these HUs are different. For the HUs with HH 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 ARs, and the percentage of ARs that cannot be evaluated is higher than observed for NRFU enumerations.

Both NRFU and ARs 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 AR address but link to a census address elsewhere causes concern that these ARs 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 AR PIKs were found, the CCM did not evaluated accuracy of the enumeration of the people at the address. Therefore, the accuracy of AR 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 ARs in both HUs with both types of respondents. Keep in mind that all the AR records have PIKs, but the Census Bureau procedure may or may not be able to assign PIKs to the census enumerations.

The assignment of PIKs to the combined CCM records proved crucial to being able to evaluate the ARs in HUs enumerated during NRFU. Therefore, the percentage of NRFU enumerations that received PIKs is an evaluation tool in and of itself. 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 system attempted to assign PIKs, 73% (SE = 0.9%) of those in HUs enumerated by proxy received PIKs while 92% (SE = 0.2%) of those enumerated by a HH 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 HH 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 HH 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.

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

For the data-defined enumerations with PIKs, 68% from proxy respondents and 91% from HH 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 HH member respondents. When the denominator for those without PIKs includes whole person imputations, the correct enumeration rate for proxy respondents is 41%. For HH 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 HH members than for proxy respondents.

Table 4. Weighted distributions of combined CCM residence status for enumerations in NRFU HUs by NRFU respondent type and PIK status (**shown in 1,000's)**

	Proxy			**HH 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 HUs in the combined CCM with several criteria for the enumerations included in the denominator by type of NRFU respondent.
(**shown in 1,000's)**

	Proxy respondent			**HH member respondent**		
Status of enumerations in denominator	**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 & imputed	3,903	1,609	41%	3,871	2,398	62%

3.2 Analysis of records for entire households

Our ultimate interest is the quality of ARs on a household basis. Our analysis examines two measures. One is the percentage of HUs where the population counts from NRFU and ARs are equal. The other is the percentage of NRFU HUs where the combined CCM determines the AR roster is perfect. These are descriptive analyses with unweighted data.

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

Table 6. Unweighted comparison of HU population counts from NRFU and ARs by respondent type

	proxy		**HH member**	
HU population counts	**HUs**	**%**	**HUs**	**%**
Same AR & census	2,685	51%	8,633	51%
Different AR & census	2,625	49%	8,243	49%
Total	5,310	100%	16,876	100%

Therefore, we examine the accuracy of the ARs on a household basis for the 5,310 proxy HUs and 16,876 HH member HUs that have ARs. Table 7 shows the percentage of HUs in the following categories as determined by the combined CCM:

- AR Perfect – All AR persons in the HU are Census Day residents at the address and no Census Day residents are omitted from the AR roster.
- AR Erroneous Enumerations and Unresolved Enumerations (E&Us) –At least one AR record in the HU 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.
- AR 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 AR roster for the address.

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

HU status	**Proxy respondents**	
AR E&U	3,180	59.9%
AR Perfect	1,722	32.4%
AR Omissions	408	7.7%
Total	5,310	100.0%

When the ARs in the 5,310 proxy HUs 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 ARs for 408 (7.7%) of the HUs 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 AR did not link to a combined CCM record with a resolved residence status.

4. Summary and Next Steps

To conclude, we return to our research questions.

1) **Are proxy responses more or less accurate than administrative records?**

Answering this question is not as straightforward as it sounds. Our investigation used the 10,416 NRFU HUs with proxy respondents and 16,876 NRFU HUs with HH member respondents in both the CCM P-sample and E-sample that the census classified as occupied. We studied the AR and census records assigned to the addresses for these HUs and used the combined CCM records to evaluate the accuracy of the records from each source.

The major findings from our study follow:

- Approximately half of the NRFU proxy HUs do not have ARs in the IRS 1040 and Medicare files for all of 2010. These two administrative sources include some information on household composition. Unless additional high-quality AR sources that cover these addresses can be found, these HUs will need contact by NRFU enumerators or whole HH imputation.
- By almost any standard, proxy enumerations that can be assigned PIKs tend to be in the correct location. Therefore, one indicator for a higher quality NRFU enumeration appears to be whether it has enough information for the Census Bureau's PVS algorithm to assign a PIK.
    - Many 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 at the correct location.
    - Whole household imputations are 20.7% of enumerations in proxy HUs and 1.4% of enumerations in HH member HUs.
- When the NRFU enumerations had enough information for the PVS system to attempt to assign a PIK, the percentages that received PIKs were 73% of the proxy enumerations and 92% of the HH member enumerations.
    - When the whole person imputations are included in the denominator, the percentages receiving PIKs are 58% for proxy enumerations and 91% for HH member.
    - Possibly enumerations could be PIK-ed as they come in during NRFU for a quality assessment.
- The combined CCM found that an unweighted 32% of the proxy HUs with ARs had perfect HH composition. That means that all the AR persons in the HU were Census Day residents and no Census Day residents were omitted from the AR roster. The enumerations with unresolved residence status were not considered to be at the correct location although some likely are but there is not enough information to make a determination.
    - Household-based matching may be able to produce additional links to combined CCM records with a resolved residence status than were found using PIK-based matching.
- When focusing only on population count, the percentage of HUs have an AR count that agrees with the census count is an unweighted 51% among HUs with proxy respondents and among HUs with HH member respondents.

    - An unweighted 34% of proxy HUs have an AR count that agrees with the census count and all the AR PIKs in combined CCM giving them the potential for evaluation and while 44% of HH member HUs meet the same criteria.
- Duplication may be a problem when using ARs to enumerate whole HHs. Unweighted, 17% of ARs in proxy HUs and 11% of ARs in HH member HUs linked to a census enumeration at an address other than the AR address. Also troubling is that 27% of ARs in proxy HUs and 12% of ARs in HH member HUs did not link anywhere in the census.
    - For ARs that link to a census enumeration and address other than the AR address,
        - Using the AR at its address may create a duplicate in the census. Census operations may need to search census enumerations, particularly self-responses, to be sure that an AR enumeration is not a duplicate. The addition of questions regarding other residences to the census questionnaire may aid in avoiding duplicates.
        - The use of household-based matching between the ARs that link to a census address other than their AR addresses and the combined CCM has the potential for finding more links. If there is a link, an examination may provide information about the reason for the person being at both addresses. If no link, then the enumeration outside sample block could be person's only enumeration.

2)How does the quality of proxy responses vary?

The next steps in our research will concentrate on investigating how the quality of the proxy responses may vary. In further investigations, we will examine the demographic, geographic, and socio-economic characteristics of the HUs where the combined CCM found their individual ARs to be perfect, that is, the exact household members were correctly enumerated versus those HUs with ARs that had errors or could not be evaluated. We also will investigate the relationship between operational characteristics, such as the number of prior contact attempts, and correct proxy responses and identify characteristics of HUs with complete correct administrative records among NRFU proxy responses. We plan to merge data from the Planning Database (U.S. Census Bureau 2015) to be able to do this investigation. The methods we plan to use in this investigation include decision trees and other multivariate statistical methodologies.

In summary, our results to date indicate that the design of NRFU operations would profit by including strategies to obtain high-quality proxy responses, possibly secondary in priority to strategies to obtain responses from HH members. Such strategies include developing contact tactics that incorporate times when knowledgeable proxy respondents are likely to be accessible, namely at home for neighbors or on the premises for multi-unit building managers. In addition, design the training of interviewers to emphasize that the name and age of the residents from proxy respondents are priorities. The main support for this recommendation is that NRFU proxy respondents who can provide a name and two characteristics appear to report high quality information and therefore, are better than census whole person imputations. Among the enumerations that were not whole person imputations, 73% of the enumerations in proxy HUs 92% for enumerations in HH member HUs could be assigned PIKs. Since the unweighted percentage of the HUs with proxy respondents did have ARs in IRS 1040 and Medicare files for all of 2010 is 51%, ARs cannot be considered a cure-all at this point in time. Unless additional high-quality sources of ARs can be found for the 49% with no ARs in these IRS and Medicare files, whole person imputations are the only other alternative for the HUs. However, using ARs strategically has the potential to save money during NRFU.

5. References

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. U.S. Census Bureau. Washington, DC.

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

Mule, T. (2012) "2010 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. U.S. Census Bureau. Washington, DC.

Mulry, M. H. and Spencer, B. D. (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. http://www.niss.org/sites/default/files/Mulry_september2012.pdf

Olson, D. and Griffin, R. (2012) "2010 Census Coverage Measurement Estimation Report: Aspects of Modeling." DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-G-10. U.S. Census Bureau. Washington, DC.

U.S. Census Bureau (2015) Planning Database. U.S. Census Bureau. Washington, DC. Last accessed 9/29/2015 at http://www.census.gov/research/data/planning_database/

Wagner, D., & Layne, M. (2014). The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) Record Linkage Software CARRA Working Paper Series. #2014-01. Last accessed 9/22/2014 at https://www.census.gov/srd/carra/CARRA_PVS_Record_Linkage.pdf