Considerations in Designing the 2020 Post-Enumeration Survey Sample[1]

Laura A. Davis[*], T. Trang Nguyen[*], and Courtney Hill[*]

[*]U.S. Census Bureau, 4600 Silver Hill Rd, Suitland, MD 20746

Abstract

This paper provides an overview of the 2020 Post-Enumeration Survey sample design. The 2020 Post-Enumeration Survey is designed to measure the coverage of the 2020 Census. It uses a complex sample design to support estimates of net coverage and components of census coverage. This paper highlights technical aspects of the design, focusing on those that have changed from a prior post-enumeration survey. It discusses the allocation of sample under multiple phases. Other topics include sample stratification given geographic changes to the primary sampling units, re-stratification occurring in each phase of the design, and critical decisions about subsampling in the final sampling phase.

Key Words: Post-Enumeration Survey, area sample, multi-phase sample design, Decennial Census, sample size calculation, stratification

1. Introduction

Every ten years, the U.S. Census Bureau conducts a census of the population and housing as mandated by the United States Constitution. Population counts from the decennial census are used for apportionment of seats of the United States House of Representatives. These counts also inform federal funding for programs in various areas of the country as well as redistricting within states. For these reasons, the accuracy of population counts resulting from a decennial census are of extreme importance. There is also a vested interest in evaluating the accuracy of the decennial census to identify areas for improvement in the following census.

According to Mulry and Cantwell (2010), the Census Bureau applies two methods to evaluate censuses: dual system estimation using a post-enumeration survey and demographic analysis. A post-enumeration survey has been used to estimate the size of the true population and census coverage starting with the 1980 Census. Dual system estimation requires matching housing units or people in an independent survey to records in the census. Demographic analysis is a method used to form national population size estimates for subgroups such as age, sex, and race using data from administrative records. The Census Bureau uses birth and death records, estimates of international migration, and data from Medicare to produce demographic analysis estimates (Devine et al., 2011). There are advantages and disadvantages to both evaluation methods as discussed in Mulry and Cantwell (2010). In our paper, we discuss the sample design of the 2020 Post-Enumeration

---

[1] Disclaimer: This paper is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau. This paper meets all of the U.S. Census Bureau's Disclosure Review Board (DRB) standards and has been assigned DRB approval number DRB-B0004-DSSD-20180924.

Survey (PES), which will be used to evaluate the 2020 Census. There will also be demographic analysis estimates produced for the 2020 Census; however, they are not discussed in this paper.

The 2020 PES includes a sample of housing units and people living in housing units in the 50 states, the District of Columbia, and Puerto Rico, excluding remote Alaska. Group quarters and people living in group quarters are also excluded from the PES estimates. The 2020 PES sample is a general-purpose probability sample conducted in three phases. In the first phase, a stratified systematic sample of primary sampling units, called Basic Collection Units (BCUs) are selected. The second phase subsamples the small BCUs with few housing units. The third phase of sampling selects segments of housing units within BCUs containing 58 or more housing units. Within the sample areas, two samples of housing units are independently formed, once from the census (called the E sample) and one from a PES area listing (called the P sample).

The same sample design used for the prior post-enumeration survey, the 2010 Census Coverage Measurement (CCM) documented in Konicki (2013) will be used for the 2020 PES. There is no plan for a sample reduction of primary sampling units[2] that was implemented in 2010. We made modifications to some of the 2010 CCM design features to improve the 2020 PES.

The primary sampling unit has changed since the last decade. In the 2010 CCM, groups of geographically contiguous 2010 census collection blocks were clustered together to form the primary sampling units. The primary sampling units for the 2020 PES are BCUs instead of block clusters. BCUs are geographic units that are smaller than tracts and replace the 2010 census collection blocks. They must follow visible features and military land boundaries, and may only contain one census type of enumeration area[3]. BCUs also respect state, county, and tract boundaries. Unlike the 2010 CCM, the 2020 PES uses the BCUs without additional clustering; therefore, they contain fewer housing units on average compared to the 2010 block clusters. Because the 2020 design is less clustered than the 2010 design, it should yield lower sampling variance, but will require us to list more BCUs than block clusters to reach a similar housing unit sample size. This change also impacts the stratification thresholds using housing unit counts in multiple phases of the design.

In this paper we discuss the technical aspects of each phase of the 2020 PES sample design and improvements to some 2010 CCM design features. The changes include:
- Stratifying and allocating the BCUs.
- Allocating the first-phase sample using state-specific averages as opposed to a national average.
- Collapsing subsampling strata in the second phase of the P sample.
- Collapsing subsampling strata for census-only housing units in the E-sample identification.

---

[2] An initiative to reduce nonsampling error in the 2010 CCM required decreasing the sample size so that the surplus funds from the reduced workload could be put toward this initiative. Entire primary sampling units were systematically reduced after area listing of housing units (Whitford, 2009).
[3] The type of enumeration area represents the predominant approach to conducting the 2020 Census in a given geographic area. The type of enumeration area assignment is based on area characteristics to maximize respondent participation. The are six census types of enumeration for 2020.

- Removing subsampling for added census housing units in the E-sample identification.

Section 2 provides an overview of the P and E samples and how they facilitate the dual system estimation. Sections 3 through 6 discuss the technical aspects of the three phases of the PES sample design.

2. P and E Samples

Dual system estimation is a method to estimate the size of a population. The difference between the dual system estimate and the census count is called net coverage. Dual system estimation is based on capture-recapture methodology and has been used by the Census Bureau since 1980. The E sample provides the capture records and the P sample provides the recapture records. The P sample is an area-based sample and the E sample contains census housing unit and person enumerations in the same sample areas as the P sample. The Cartesian product of the P- and E-sample records can be placed into one of four cells shown in Figure 1.

In P Sample

		Yes	No	Total
	Yes	N_{11}	N_{12}	N_{1+}
In E Sample	**No**	N_{21}	N_{22}	N_{2+}
	Total	N_{+1}	N_{+2}	N

Figure 1: Classification of P and E Sample Records into 2x2 Matrix

We use the standard Petersen (1896) or Sekar-Deming dual system estimator to measure the size of the true population. Assuming that the P and E samples are independent and that each unit has the same chance of being in the E sample and the P sample,

$$\widehat{N} = N_{1+}\frac{N_{+1}}{N_{11}}.$$

(1)

In Equation 1, $\widehat{N}$ is an estimator for the unknown population total N. The weighted E-sample total (N_{1+}), the weighted P-sample total (N_{+1}), and the weighted number of records in both samples (N_{11}) are observed. As discussed in Mulry and Cantwell (2010) and Mule (2008), Equation 1 is adjusted to exclude erroneous enumerations and records that have insufficient information needed for matching.

3. Phase I

The first phase identifies an area sample of BCUs for the P and E samples. A stratified systematic sample of BCUs is selected to meet the P-sample housing unit interview sample size goal of 180,000 in the 50 states, District of Columbia, and Puerto Rico.

3.1 BCU Stratification
The 2020 PES design stratifies BCUs by the cross-classification of
- State[4] (52 categories),
- Size (3 categories),
- Tenure status (2 categories), and
- American Indian Reservation (AIR) status.

Within state, the first level of classification categorizes BCUs by size into three mutually exclusive categories based on their expected housing unit counts from the most recent Census Bureau Master Address File as follows:
- Small BCUs with 0 to 2 housing units.
- Medium BCUs with 3 to 57 housing units.
- Large BCUs with 58 or more housing units.

The second classification categorizes medium and large BCUs into non-owner and owner strata based on percent of non-owner households in a BCU. Non-owner strata contain BCUs with 40 percent or more of non-owner households based on 2010 Census data and owner strata contain BCUs with less than 40 percent of non-owner households.

For 26 states with sufficient American Indian population living on AIR in 2010, an AIR stratum is defined to ensure sufficient precision of coverage estimates for the AIR domain. This stratum contains BCUs located on AIR or associated trust land that have 3 or more housing units. There may be AIR in other states, but with insufficient American Indian population to support a separate AIR stratum.

---

[4] State from this point forward includes the 50 states, District of Columbia, and Puerto Rico unless otherwise specified.

Table 1 summarizes the six mutually exclusive sampling strata within each state.

Table 1: Strata for Selecting Basic Collection Unit Sample for Area Listing

Phase I Stratum	Size Definition	Tenure Definition
(1) Small	0-2 housing units	Not applicable
(2) Medium Owner	3-57 housing units	less than 40% of non-owner households
(3) Medium Non-Owner	3-57 housing units	40% or more of non-owner households
(4) Large Owner	58+ housing units	less than 40% of non-owner households
(5) Large Non-Owner	58+ housing units	40% or more of non-owner households
(6) AIR (26 states only)	3+ housing units	Not applicable

AIR stands for American Indian Reservation.

We oversample large BCUs in Phase I, but then subsample housing units from the large BCUs in the third phase. Since there is intra-cluster correlation for housing units within the same BCU, this design allows us to reduce the design effect of the sample. For a fixed housing unit sample size, we can visit more BCUs, thereby reducing the clustering effect in our sample design. Given the fact that households in a BCU are expected to have similar characteristics, there is little new information gained by interviewing all of them.

We also oversample BCUs with 40 percent or more non-owner households to help with the reliability of coverage estimates for the non-owner population, which historically has been undercounted in the census (Mule, 2012). Moreover, our research found that the non-owner stratum included a higher percentage of minority race groups than the owner stratum, which will help improve the reliability of the minority population estimates.

The stratum definitions in Table 1 are similar to those of 2010 CCM except the size definitions for medium and large strata, which changed to reflect the difference between the size of the BCUs and block clusters. In 2010, the medium stratum contained block clusters with 3 to 79 housing units and the large stratum contained block clusters with 80 or more housing units.

We studied two stratifications for medium and large BCUs: the 2020 size definition in Table 1 and the 2010 size definition. The goal of the 2020 definition was to preserve the same proportion of frame housing units in medium and large strata from the 2010 design. On average, block clusters are almost 1.5 times larger than BCUs in terms of housing units. The BCU frame count is higher than the block cluster count, with approximately 5.1 million and 3.9 million, respectively.

Table 2 shows distributions for three different size definitions:
- 2010 Definition – Block Cluster Frame. This stratification uses the cut-offs of the 2010 size variable applied to the 2010 block cluster frame.

- 2010 Definition – BCU Frame. This stratification uses the cut-offs of the 2010 size variable applied to the 2020 BCU frame.
- 2020 Definition – BCU Frame. This stratification uses the cut-offs of the 2020 size variable applied to the 2020 BCU frame. This stratification was designed so that the distribution of housing units in each category matches the distribution of housing units in each size category from 2010.

Table 2: 2010 and 2020 Size Definition Frame Results by Strata

	Small	Medium	Large	AIR
Percent Block Clusters or BCUs				
2010 Definition - Block Cluster Frame	21.4	69.2	9.0	0.4
2010 Definition - BCU Frame	13.5	80.1	5.9	0.5
2020 Definition - BCU Frame	13.5	76.8	9.2	0.5
Percent Housing Units				
2010 Definition - Block Cluster Frame	0.3	51.5	47.8	0.4
2010 Definition - BCU Frame	0.2	60.0	39.5	0.3
2020 Definition - BCU Frame	0.2	51.9	47.6	0.3

AIR stands for American Indian Reservation.
BCU stands for Basic Collection Unit.
Source: U.S. Census Bureau, 2010 Census Coverage Measurement, Master Address File, and Geography Files.

The percent of BCUs in the large stratum under the 2020 size definition is similar to the percent of block clusters in the large stratum (9.2 percent to 9.0 percent, respectively), but the percent of medium BCUs is higher than the percent of the medium block clusters (76.8 percent to 69.2 percent, respectively). This difference is caused by our decision to keep the definition for the small stratum the same as the 2010 CCM with 0 to 2 housing units. Since the small stratum contributes a very small percentage of housing units to the total, there were no strong justifications for changing the upper bound.

By simply changing the size definition for the stratification without changing the allocation method, we found the housing unit area listing workload estimate for 2020 to be similar to the target workload from 2010. We found that when applying the 2010 size definition to the BCU frame yielded an unrealistically higher housing unit area listing workload estimate for the 2020 PES than the target.

3.2 Sample Allocation to Stratum
We will need to list approximately 10,500 BCUs[5] (7,800 medium and large BCUs, 2,200 small BCUs, and 500 AIR BCUs) to support 180,000 housing unit interviews targeted for the P sample.

The 2,200 small BCU sample size was determined by multiplying the sample size to frame size ratio for the small block cluster sample in the 2010 CCM design to the small BCU count on the frame. Based on our research, it should provide a sufficient number of sample BCUs for subsampling in the second phase. The allocation of the 2,200 small BCUs to each state (excluding Puerto Rico) is proportional to the number of small BCUs in each state. Puerto Rico has a fixed number (120) of small BCUs assigned.

---

[5] Included in this number is approximately 400 BCUs for Puerto Rico (280 medium and large BCUs; and 120 small BCUs). There are no AIR BCUs in Puerto Rico.

The 500 AIR BCU sample size was calculated to support a target AIR housing unit interview sample size of 10,500. First, the housing unit sample size was proportionally allocated to the 26 states assigned that stratum based on the 2010 Census population count of American Indians living on AIR in each of those states. Then the BCU sample size was determined for each of those states using the expected average number of housing unit interviews per BCU in that state.

The 7,800 medium and large BCU sample size was determined by starting with the target housing unit interview sample size and using an expected average number of interviews per BCU by state to obtain the total number of BCUs we will need to list. We then allocate those BCUs to the four medium and large strata within the states (medium owner, medium non-owner, large owner, and large non-owner). The remainder of this section describes the allocation process in detail.

The first step for the medium and large BCUs is allocating approximately 159,500 of the 180,000 housing unit interview sample size to the 50 states and District of Columbia proportional to the population in each state. This gives us the housing unit interview sample size for state s, h_s. The exceptions to this rule include slight increases in the sample for (1) small states to meet the minimum housing unit interview sample size requirement and (2) Hawaii to help the reliability of the Native Hawaiian or Pacific Islander estimates. Additionally, Puerto Rico has a predetermined housing unit interview sample size of 7,500 that does not need to be allocated to lower geography. Note that the remaining 13,000 housing unit interview sample size is the number of housing units we expect to come from the small and AIR BCUs that do not need to be allocated using the method described below.

For each state s, we then calculate the medium and large BCU sample size, TB_s, as

$$TB_s = \frac{h_s}{\bar{H}_s} \tag{2}$$

where
- h_s is the housing unit interview sample size in state s and
- $\bar{H}_s$ is the expected average number of housing unit interviews per BCU for state s.

Next, we assign TB_s to medium owner, medium non-owner, large owner, and large non-owner strata to compute the BCU sample size, $b_{s,k}$, for each state s and stratum k. During this step we incorporate two oversampling objectives.

To improve the precision of minority and non-owner estimates, we oversample BCUs with higher proportions of non-owner households at a rate 1.5 times higher than BCUs categorized as owner.

Also, the large BCUs are selected with relatively higher probability than medium BCUs to prepare for the subsampling of housing units within these BCUs in Phase III. Phase III subsampling brings the overall probability of selection for housing units in large BCUs more in line with the overall probability of selection for housing units in medium BCUs. One of our design objectives is to minimize the amount of weight variation in the sample. This will also reduce the disproportionate influence of higher weighted housing units on point estimates and standard errors. Finally, oversampling large BCUs increases the number of unique BCUs in sample to provide a more geographically diverse housing unit sample.

A portion of TB_s is allocated first to two medium strata, medium owner and medium non-owner, proportional to the number of housing units expected in the BCUs on the frame within state s in owner or non-owner category g using

$$m_{s,g} = M_{s,g} \frac{h_{s,g}}{H_{s,g}} \tag{3}$$

where
- $m_{s,g}$ is the medium BCU sample size in state s and category g,
- $M_{s,g}$ is the medium BCU frame count in state s and category g,
- $H_{s,g}$ is the total number of housing units on the frame in state s and category g,
- $h_{s,g}$ is the housing unit interview sample size accounting for oversampling of non-owner in state s and category g, and
- $g = 1$ for owner and 2 for non-owner.

The housing unit interview sample size for state s and category g, $h_{s,g}$, is calculated as

$$h_{s,g} = h_s \frac{D_g H_{s,g}}{\sum_{g=1}^{2} D_g H_{s,g}} \tag{4}$$

where D_g values are the same for all states and defined as
- $D_g = 1$ for $g = 1$ (owner) and
- $D_g = 1.5$ for $g = 2$ (non-owner).

Finally, the remaining TB_s is allocated to two large strata, large owner and large non-owner, as

$$l_{s,g} = \left(TB_s - \sum_{g=1}^{2} m_{s,g}\right) \frac{L_{s,g} \frac{h_{s,g}}{H_{s,g}}}{\sum_{g=1}^{2} L_{s,g} \frac{h_{s,g}}{H_{s,g}}} \tag{5}$$

where
- $l_{s,g}$ is the large BCU sample size in state s and category g and
- $L_{s,g}$ is the frame count in state s and category g.

Note that the only exception to our allocation methods was a minimum sample size requirement of two BCUs from each stratum so that sampling errors can be estimated for all strata independently.

This allocation method controls the sample weight relationships among the strata. The differential sampling factor of 1.5 can be seen in the ratio of weights between the owner and non-owner category by size within state. Also, this calculation ensures that the total number of medium and large BCUs are constrained to the TB_s (the overall medium and large BCU sample size in each state), which is based on $\overline{H}_s$ (an expected average housing unit interviews per BCU in each state).

We evaluated two different values of the expected average housing unit interviews per BCU for each state and their impact on the overall medium and large BCU sample size in each state. The first expected average was calculated nationally across all states, which was the method used during the 2010 CCM. In our research, this national expected average

number of housing unit interviews per BCU was 22. The second method calculated an expected average for each state. These values ranged from 16 to 30 and tended to be smaller in small population states. We found that by using state specific expected averages, the BCU allocation is more efficient compared to using a national average for all states. The BCU sample size needed to support the housing unit interview sample size was reduced in higher population states where we do not need the extra BCU sample, but increased in lower population states where the extra BCU sample is needed to support an oversample of large BCUs.

3.3 BCU Sample Selection

The BCUs are first sorted within each stratum. Our sorting variables, which defines the ordering for the systematic sample, differs depending on the stratum. All BCUs are sorted on minority status and geography codes, but the size sort is not relevant for the small BCUs. The sort is nested and based on the following characteristics:

- Size category based on the expected housing unit count for each BCU from the most recent Census Bureau Master Address File (seven categories: $3 - 9$, $10 - 26$, $27 - 57$, $58 - 101$, $102 - 197$, $198 - 499$, and 500 or more housing units). A goal in creating these categories is to have approximately the same number of housing unit in each group.

- Minority status based on percent of minority population[6] in a BCU:
  - Minority BCU has more than 35 percent of minority population.
  - Non-minority BCU has 35 percent or less of minority population.

- Geography codes (county then BCU).

The sorting provides an implicit stratification, which yields a sample with a similar distribution to the frame.

The sampling interval, *SI*, for each stratum *k* within each state *s* is calculated as

$$SI_{s,k} = \frac{B_{s,k}}{b_{s,k}} \tag{6}$$

where $B_{s,k}$ is the BCU frame count for a state and stratum and $b_{s,k}$ is the BCU sample size.

4. Phase II

The second sampling phase selects a subsample of BCUs from the Phase I small sampling stratum after an area listing of housing units. We use a double-sampling technique by selecting a larger sample of small BCUs for area listing and then subsampling using updated housing unit counts (Kish, 1965). This is done to reduce a BCU's influence on the estimates when more housing units are found than expected. Additionally, small BCU subsampling reduces travel cost by reducing the number of BCUs that have few housing units. Conducting followup interviews in small BCUs is more expensive per housing unit than in medium or large BCUs.

---

[6] A minority person is based on the race and Hispanic origin of the people counted in 2010 Census. All BCUs in Puerto Rico are classified as minority.

We re-stratify the Phase I small BCUs into four strata within each state based on the updated housing unit counts and American Indian Country[7] status (see Table 3). This is a change from the 2010 CCM design. For that survey, we re-stratified the small block clusters into five strata. For the 2020 PES, we will combine the 3 to 5 housing unit stratum and the 6 to 9 housing unit stratum to create one stratum with 3 to 9 housing units. This collapsing was necessary because the number of block clusters within each stratum observed in 2010 was extremely small. Approximately 60 percent of the 3 to 5 and 6 to 9 housing unit strata either (1) did not contain enough block clusters to subsample (making them certainty strata) or (2) contained no block clusters to subsample.

Table 3: Subsampling Strata Within State for Selecting Small Basic Collection Units

Phase II Stratum	Size Definition	On American Indian Country?
1	0-2 housing units	No
2	3-9 housing units	No
3	10+ housing units	Not applicable
4	0-9 housing units	Yes

The strata were formed so that BCUs with more housing units than we initially expected will be selected at a higher probability to keep the overall sampling weights lower. BCUs in stratum 3 are retained in sample with certainty. To control variability for American Indian population estimates, we also retain all BCUs in stratum 4 in sample.

The national target percent retain for the 50 states and District of Columbia for each stratum j, where $j = 1$ for stratum 1 and $j = 2$ for stratum 2 is

$$R_j = \begin{cases} 10\%, if\ the\ BCU\ is\ in\ Phase\ II\ stratum\ 1 \\ 30\%, if\ the\ BCU\ is\ in\ Phase\ II\ stratum\ 2 \end{cases}$$

While these national percentages are similar to those of the 2010 CCM, the 2020 PES calculates a second-phase BCU sample size for each stratum across all states, then allocates that to individual states after the updated housing unit counts are available. The 2010 CCM used a fixed sampling rate within each stratum across all states, whereas the 2020 PES sampling rates for strata 1 and 2 will differ from state to state.

The national target BCU sample size for each stratum j, stb_j, is calculated as

$$stb_j = R_j \times STB_j \tag{7}$$

where STB_j is the national BCU frame count for stratum j. The sample size stb_j is then allocated to each state s proportional to the state updated housing unit count among small BCUs after area listing. For Puerto Rico, the target percent retain for strata 1 and 2 is also 10 and 30, respectively.

An equal probability sample of BCUs is selected systematically within each state s and stratum j, after sorting on minority status and geography codes as defined in Section 3. Similar to Phase I, we want to select a minimum of two BCUs per stratum. Since the frame size is very small and the strata are formed within each state, we are expecting that there will be strata with zero or one BCUs. When there is one BCU in a stratum, that BCU is

---

[7] This includes both on AIR and on Indian Country but off AIR.

retained in sample with certainty. These situations do not contribute to the second-phase sampling variance.

Overall, based on our 2020 PES sample design research, we are expecting to retain approximately 22 percent of the 2,200 small BCUs from Phase I in sample.

5. Phase III P-Sample Selection

The P-sample portion of the third phase identifies housing units from the sample BCUs for inclusion in the P sample. These housing units are interviewed to collect rosters of people in the housing units as well as other information needed for matching to census records. We expect to list approximately 570,000 housing units and retain less than a third in sample for the 2020 PES.

For BCUs with 57 or fewer housing units observed from area listing, all area listed housing units are included in the P sample. For BCUs with 58 or more housing units observed from area listing, we form groups of contiguous area listed housing units, called segments, and select one or more segments from each BCU for inclusion in the P sample. Recall that this subsampling was planned for in Phase I by oversampling large BCUs with 58 or more expected housing units. The segments in a BCU have an approximately equal number of housing units. A systematic sample of segments is selected with equal probability within each stratum defined in Section 5.1. A geographic sort is used to order segments for the systematic sampling. All area listed housing units in selected segments are included in the P sample and eligible for interviews. Selecting segments of housing units allows geographically compact interviewing and facilitates the overlap of the P sample with the E sample.

5.1 BCU Re-Stratification
In the first step of identifying the P-sample housing units, all medium and large sample BCUs selected in the first phase and small BCUs retained from the second phase are categorized into nine strata within each state (see Table 4). The BCUs are re-stratified by the first-phase characteristics: AIR status, size (small, medium, and large), and tenure status (owner and non-owner). The BCUs are additionally stratified by their area listed housing unit count.

Table 4: Strata for Within-BCU Housing Unit Subsampling

Phase III Stratum	BCU Size from Phase I	Tenure Status	Size Definition[1]	Subsampling Stratum?
1	Small[2]	Not applicable	< 58 housing units	No
2	Medium or Large	Owner	< 58 housing units	No
3	Medium or Large	Non-owner	< 58 housing units	No
4	Large	Owner	58+ housing units	Yes
5	Large	Non-owner	58+ housing units	Yes
6	Small[2]	Not applicable	58+ housing units	Yes
7	Medium	Owner	58+ housing units	Yes
8	Medium	Non-owner	58+ housing units	Yes
9	AIR (26 states only)	Not applicable	0+ housing units	No

BCU stands for Basic Collection Unit.
AIR stands for American Indian Reservation.
1 This size definition is based on the number of area listed housing units.
2 Stratum 1 also includes Phase I small BCUs on American Indian Country even if they have 58 or more area listed housing units because these BCUs are not eligible for subsampling. These BCUs are therefore not included in Stratum 6.

The sampling frame for the P sample in the third phase is the list of valid housing unit addresses obtained from area listing that are in the sample BCUs. For BCUs in the non-subsampling strata 1, 2, 3, and 9, all housing units are retained in the P sample with certainty. Strata 4 through 8 contain BCUs that were in the first-phase small, medium, and large strata and have 58 or more housing units observed from area listing. Housing units in these BCUs are eligible for subsampling.

5.2 Sample Allocation to Stratum
Recall in the first phase, we determined the target housing unit interview sample size for the medium and large strata for each state s, h_s, by allocating approximately 159,500 housing units to each state (excluding Puerto Rico) proportional to the population in each state.

In the third phase, we allocate h_s to each subsampling stratum (i = 4 through 8 from Table 4) within state s using Equation 8. We first remove the total number of area listed housing units from the non-subsampling strata i = 2 and 3 in state s (i.e., $\sum_{i=2}^{3} LHU_{s,i}$) from h_s since housing units in these two strata are sampled with certainty. The remainder is allocated to the five subsampling strata proportional to the weighted number of area listed housing units in each state. Using this allocation helps achieve the target P-sample housing unit interview sample size for the state while preserving the oversampling of the first and second phases. The target housing unit interview sample size for each subsampling stratum i in state s, $t_{s,i}$, is calculated as

$$t_{s,i} = (h_s - \sum_{i=2}^{3} LHU_{s,i}) \frac{(Weight_{s,i})(D_i)(LHU_{s,i})}{\sum_{i=4}^{8}(Weight_{s,i})(D_i)(LHU_{s,i})} \tag{8}$$

where

- $Weight_{s,i}$ is the product of the first and second-phase BCU sampling weights for each subsampling stratum i in state s. This weight is the same for all BCUs in subsampling stratum i in state s.
- D_i is the first-phase differential sampling factor applied to the subsampling stratum i based on tenure status (i.e., $D_i = 1.5$ for $i = 5$ or 8 and $D_i = 1$ for $i = 4, 6$, or 7). Note that D_i is the same for all states.

The within BCU sampling rate for each subsampling stratum i in state s, $WBSR_{s,i}$, is calculated as

$$WBSR_{s,i} = \frac{t_{s,i}}{LHU_{s,i}}.$$

(9)

For non-subsampling strata ($i = 1, 2, 3$, and 9), $WBSR_{s,i}$ is set to 1. We use the $WBSR_{s,i}$ when calculating the number of segments to form in each BCU.

5.3 Segment Formation

Selecting a subsample of housing units requires forming segments of contiguous area listed housing units. In a Phase III stratum i ($i = 1$ through 9 from Table 4) in a state s, a fixed number of segments, $NOS_{s,i}$, is formed within each BCU. This method yields similar-sized segments within a BCU, but different-sized segments across BCUs within the same stratum. The $NOS_{s,i}$ within a BCU is a function of the $WBSR_{s,i}$ and is calculated as

$$NOS_{s,i} = \begin{cases} \frac{1}{WBSR_{s,i}}, & if \ WBSR_{s,i} \leq 0.5 \ or \ WBSR_{s,i} = 1 \\ \frac{1}{(1-WBSR_{s,i})}, & if \ 0.5 < WBSR_{s,i} < 1 \end{cases}.$$

(10)

We found that when $WBSR_{s,i}$ is greater than 0.5 but less than 1, the first formula in Equation 10 yields only two segments per BCU, resulting in increased sample size variation with this larger segment size. To better control our sample size when the $WBSR_{s,i}$ is between 0.5 and 1 (excluding the endpoints), the $NOS_{s,i}$ is calculated using the second formula in Equation 10.

Within a stratum i in a state s, the area listed housing units in a BCU are ordered by their adjacency on the ground based on GPS coordinates collected in the field and clerically reviewed. This puts them in geographic order around the BCU. The area listed housing units are then divided into segments based on the $NOS_{s,i}$. For example, if there are 90 area listed housing units in a BCU that have been sorted, and the $NOS_{s,i}$ is equal to three, then three segments are formed for this BCU with the first 30 housing units assigned to first segment, the next 30 housing units assigned to the second segment, and the last 30 housing units assigned to the third segment.

An equal probability sample of segments is selected systematically within each stratum i in state s with the sampling interval, $SI_{s,i}$, calculated as the inverse of $WBSR_{s,i}$. With this design, every BCUs will have at least one segment selected.

6. Phase III E-Sample Selection

The E sample contains census housing units in the same sample BCUs as the P sample and provides capture data for the capture-recapture methodology. As required for dual-system estimation, the P-sample housing units are matched to the housing units in the E sample. This is done at two separate times, the initial matching uses the initial census list and later the final matching is conducted using the final census list. The final census list contains the final inventory of census housing unit addresses. The final census list might include some housing units that were on the initial census list and might include additional housing units that were not on the initial census list.

For ease of accounting for the E-sample housing units during sample identification, we partition the sampling frame (i.e., final census list) for the E sample into three mutually exclusive parts as follows:

- Housing units in the initial census list that match with area listed housing units (i.e., matched census housing units).
- Housing units in the initial census list that do not match to area listed housing units (i.e., census-only housing units).
- Housing units not on the initial census list, but later added to the final census list (i.e., added census housing units).

The first two types of housing units (i.e., matched and census-only housing units) are identified as potentially in the E sample early on in the 2020 PES process. Identifying census-only housing units earlier allows us to interview a subsample of these units to confirm previously reported information at the same time as the interviews for the area listed housing units to cut down travel costs.

6.1 E-Sample Housing Unit Identification
A goal of the E-sample housing unit identification is to maintain geographic overlap of the P and E samples to the greatest extent possible. This overlap reduces the E sample person followup workload. This is done by mapping the final set of housing units from the final census list onto the sample BCU segments from the P sample. The E sample is comprised of matched, census-only, and added housing units. Similar to the 2010 CCM design, matched and census-only housing units are eligible of subsampling. However, there is no subsampling of added census housing units in the 2020 PES, this is a change from 2010 and discussed in Section 6.2.

Matched census housing units are also area listed housing units and are therefore subsampled as P-sample housing units during Phase III as discussed in Section 5. These matched census housing units are in the E sample if they are on the final census list.

Census-only housing units are on the initial census list, but could have been missed by area listing or erroneously enumerated in the census. These housing units need to be interviewed to confirm their census day resident information. Rather than interviewing all census-only housing units, we select a subsample of census-only housing units from the initial census list at the same time as P-sample housing unit identification. The selected census-only housing units are in the E sample if they are on the final census list.

To subsample census-only housing units, these housing units are first placed into P-sample segments. Census-only and area listed housing units are sorted together geographically

within each sample BCU. Each census-only housing unit is assigned to the P-sample segment based on geographical closeness to the area listed housing units. We assign all census-only housing units to a single segment if the BCU has zero area listed housing units.

We make the decision to subsample based on the number of census-only housing units in sample segment(s) for each BCU. No subsampling is needed if the number of census-only housing units in sample segment(s) in the BCU is 57 or less or the BCU is on American Indian Country. Otherwise, we subsample census-only housing units at a rate of 1-in-3. Within each BCU, census-only housing units from sample segment(s) are first ordered by their adjacency on the ground, then an equal probability sample of census-only housing units is selected systematically.

Added census housing units are first assigned to P-sample segments (sample and non-sample) to determine which added housing units are eligible for the E sample. All three types of census housing units on the final census list are ordered by their adjacency on the ground within each sample BCU. Then each added census housing unit is assigned to the P-sample segment of the preceding matched or census-only housing unit. We assign all added census housing units to a single segment if a sample BCU contains no matched or census-only housing units.

Once all the added census housing units within a BCU are assigned to the P-sample segments, all added census housing units in the segments selected for the P sample are identified for the E sample and followup interviews are conducted to confirm their status. There is no subsampling of added census housing units in the 2020 PES within these sample segments.

6.2 Changes to Identification of Census-Only and Added Census Housing Units

For 2010 CCM, we subsampled matched, census-only, and added census housing units. Very few segments qualified for census-only and added housing unit subsampling in 2010; therefore, we evaluated whether subsampling of census-only and added housing units is still needed for the 2020 PES. We first looked at the subsampling results from the 2010 CCM to quantify the increase to our followup workload if we kept all census-only and added census housing units in the E sample.

Table 5: Frame and Sample Size of Census-Only and Added Census Housing Units[1]

Housing Unit Type	Frame Count	Sample Count	Frame Minus Sample Count
Census-Only Housing Units	10,500	8,100	2,400
Added Census Housing Units	16,000	13,500	2,500

1 The counts in this table have been rounded.
Source: U.S. Census Bureau, 2010 Census Coverage Measurement.

The followup workloads of census-only and added census housing units are expected to increase by approximately 2,400 housing units and 2,500 housing units, respectively (see Table 5). This increase was deemed acceptable; however, further analysis justified us keeping all added census housing units in sample with certainty, but continue to subsample census-only housing units. We found that approximately 46 percent of the weighted added census housing units were determined to be erroneous enumerations. Less than 1 percent of the weighted matched and census-only housing units were found to be erroneously enumerated. We considered these two percentages to be practically different. Estimating

the number of erroneous enumerations in the 2020 Census is a primary goal of the 2020 PES. Therefore, based on these results and other considerations, we decided not to subsample added census housing units for the 2020 PES to keep as many of these housing units in the sample as possible. We did not see the same benefits to keeping all census-only housing units in sample. Therefore, we decided to subsample census-only housing units to not increase the followup workload.

In the 2010 CCM, census-only housing units were separated into five strata, within each state. Two of these strata were sampled with certainty and three were eligible for subsampling with fixed sampling rate for each stratum. To be eligible for subsampling, a block cluster (1) had to contain 80 or more census-only housing units in sampled segments and (2) could not be located in American Indian Country. The results of the 2010 CCM subsampling in the three subsampling-eligible strata across the 50 states and District of Columbia is summarized in Table 6.

Table 6: 2010 Census Coverage Measurement Census-Only Housing Unit Subsampling Results and Rates by Subsampling-Eligible Strata for All States and District of Columbia

Subsampling-Eligible Stratum	80-119 HUs	120-159 HUs	160+ Hus	Total	Percent of Full Frame[1]
Subsampling Rate	1-in-2	1-in-3	1-in-4		
Before Subsampling					
Sample Segment Count	11	9	9	29	0.4
HU Count	1,100	1,100	1,700	3,900	37.1
After Subsampling					
HU Count	550	400	400	1,350	NA
Percent HUs Retained	50.0	36.4	23.5	34.6	NA

HU stands for census-only housing unit.
NA stands for not applicable.
1 The percent calculated are those of the full frame including counts from two strata not eligible for subsampling. The denominator of the sample segment count is 6,680 segments (i.e., 0.4 percent is calculated by taking 29 divided by 6,680). The denominator of the census-only housing unit count is 10,500 housing units.
Source: U.S. Census Bureau, 2010 Census Coverage Measurement.

Census-only housing units were subsampled in only 0.4 percent of all sample segments (see Table 6). For this reason, we collapsed the number of subsampling-eligible strata within each state from three to one for the 2020 PES to reduce the number of extremely small strata, with a subsampling rate of 1-in-3. This subsampling rate should retain approximately the same percent of census-only housing units in sample as the 2010 CCM (i.e., 2010 CCM retained 34.6 percent of all census-only housing units in sample (see Table 6)).

7. Summary

This paper has provided an overview of the sample design for the 2020 PES and highlighted the major sample design changes to the methodology from the prior post-enumeration survey. These changes include

- Changing the primary sampling units from block clusters to BCUs.
- Using a state-specific average number of housing units in calculating the BCU sample size in each state instead of a national average.
- Reducing the number of subsampling strata for small BCUs during Phase II and for census-only housing units in the E-sample identification.
- Eliminating the subsampling of added census housing units for the E sample.

With these changes in place, the 2020 PES sample design should be more efficient than the prior post-enumeration survey's design.

While we are officially done researching improvements for the 2020 PES, we know what we would like to research for the 2030 post-enumeration survey. The 2020 Census will implement operational innovations such as in-office address canvassing and internet self-response. Because these operations are new, there is no prior data for us to use to evaluate the impact of these operations on census coverage and incorporate them into our 2020 PES sample design. However, once these operations have completed, if they remain for the 2030 Census, we will have data from the 2020 Census to use in designing the 2030 post-enumeration survey sample. We will also have coverage estimates from the 2020 PES to consider when designing the 2030 post-enumeration survey sample.

We have also done some research on combining different variables that are correlated with census coverage into one hard-to-enumerate score as a potential stratification variable (Heim, 2018). In the next decade we hope to incorporate the new 2020 Census operational data in the hard-to-enumerate score creation and further improve this score enough to incorporate it into the 2030 post-enumeration survey sample design.

Acknowledgements

References

Devine, J., Sink L., DeSalvo B., and Cortes R. (2011). "The Use of Vital Statistics in the 2010 Demographic Analysis Estimates." Population Division Working Paper Number 88 (POP-WP088), Retrieved from https://www.census.gov/library/working-papers/2011/demo/POP-twps0088.html.

Heim, K. (2018). "Creating a Hard-to-Enumerate Score to Stratify the 2020 Post-Enumeration Survey Sample." In Proceedings of the 2018 Joint Statistical Meetings, American Statistical Association (forthcoming).

Kish, L. (1965). *Survey Sampling*, Wiley and Sons, New York.

Konicki, Scott (2013). "2010 Census Coverage Measurement Estimation Methods: Sample Design." DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-J-01.

Mulry, M. and Cantwell, P. (2010). "Overview of the 2010 Census Coverage Measurement Program and Its Evaluations." CHANCE, 23:3, 46-51, DOI: 10.1080/09332480.2010.10739823.

Mule, T. (2008). "2010 Census Coverage Measurement Estimation Methodology." DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-E-18, Retrieved from https://www.census.gov/coverage_measurement/pdfs/2010-E-18.pdf.

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, Retrieved from https://www.census.gov/coverage_measurement/pdfs/g01.pdf.

Petersen, C.G.J. (1896). The Yearly Immigration of Young Plaice into the Limfjord from the German Sea. Report of the Danish Biological Station. 6, 1-48.

Whitford, David C. (2009). "Recommendation to Reduce Nonsampling Error in the 2010 Census Coverage Measurement Program." DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-A-43.