Creating a Hard-To-Enumerate Score to Stratify the 2020 Post-Enumeration Survey Sample[1]

Krista Heim*, Courtney Hill*, T. Trang Nguyen*, Timothy Kennel*
*Decennial Statistical Studies Division, U.S. Census Bureau, Washington, DC, 20233

Abstract
This research identifies and explores characteristics highly correlated with coverage issues for possible use in the sample design of the 2020 Post-Enumeration Survey. If the relationship is sufficiently strong, grouping primary sampling units into strata based on these variables could produce smaller variances than if the stratification had not used these variables. We look at several data sources, including the American Community Survey and extracts from the Master Address File. We create various versions of what we refer to as the Hard-to-Enumerate Score, including straightforward summarizations and model-based approaches. Further research compares the stratification of primary sampling units based on these scores to the stratification based on the prior post-enumeration survey, the 2010 Census Coverage Measurement Survey. We found that the stratification for the 2010 Census Coverage Measurement Survey resulted in similar variances to the alternative stratifications.

Key Words: coverage error models, census, post-enumeration survey, stratification, Planning Database

1. Introduction

The purpose of the 2020 Post-Enumeration Survey (PES) is to measure the coverage of the 2020 Census. The coverage measures resulting from this survey allow us to evaluate the quality of census counts as well as help improve future census processes. Past observations have shown that in a decennial census, not owner occupants and minorities, among other demographic domains, have historically had coverage issues (Schellhamer, 2010). The prior post-enumeration survey, 2010 Census Coverage Measurement (CCM) documented in Moldoff (2008), may have not fully taken these characteristics into account due to the lack of current small area data needed to identify areas with high concentrations of hard-to-enumerate people and housing units. With the availability of American Community Survey (ACS) data, in addition to other data sources, there is potential to better identify hard-to-enumerate areas and incorporate them into the 2020 PES sample design.

There are several examples of combining multiple predictors of hard-to-enumerate areas into scores or indices. A Hard-to-Count Index was used in the 2001 and 2011 UK Census. In 2001, this metric was simply a sum of the proportions of the variables such as unemployed persons, imputed households, persons whose country of birth is non-English speaking, households in multiply-occupied buildings, and households which were privately

[1] This paper is released to inform interested parties of research and to encourage discussion. The views expressed are those of the author 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 CBDRB-FY18-500.

rented (ONS, 2000). The proportions were then converted into three categories by dividing them into a 40 percent, 40 percent, 20 percent distribution at a national level, with each group assigned a value of 1 (for easiest to count), 2 (for modestly hard to count), or 3 (for hardest to count). In 2011, instead of using the proportions directly, the UK Census used predicted values from a logistic regression model to create their Hard-to-Count index (ONS, 2009).

At the U.S. Census Bureau, Bruce et al. (2001) developed a Hard-to-Count Score for identifying areas that were likely to have high census nonresponse using the Census Bureau 2000 Planning Database. This score was created by using variables including age, employment status, educational attainment, and occupancy status that were highly correlated with nonresponse. The variables were sorted individually from high to low across tracts and assigned 12 ranks based on specific percentiles. These ranks were then summed over the variables to create the Hard-to-Count Score. More recently, Erdman and Bates (2017) created a model-based Hard-to-Count Score to stratify areas by their propensity to self-respond in sample surveys and censuses. The model was selected based on "The Census Return Rate Challenge" that was open to the public and asked participants to model 2010 Census mail return rates using variables in the Census Bureau 2012 Planning Database. Using the results of this challenge, their final modified model created a Low Response Score (LRS), summarizing twenty-five variables that were highly predictive of mail response at the block-group level. The single most influential predictor in the winning model was the percentage of renter households. Previous research noted wide variation in census participation between homeowners and renters as far back as the 1990 Census.

In this research, we analyzed new potential stratification variables that had not been used in prior post-enumeration surveys. Variables that were highly correlated with certain coverage statuses (i.e., erroneously enumerated records in the census and records that were in the PES but not found by the census) are good candidates for the 2020 PES to use in designing the sample to target areas where the census may have had difficulty. If the correlation was sufficiently high, grouping primary sampling units into strata based on these variables could have produced smaller variances than if the stratification had not used these variables.

Based on past research of constructing scores, this paper explains how we calculated and compared two versions of a Hard-to-Enumerate (HTE) score. The first score was a standardized average approach that resembled the 2001 UK score which used simple summaries of variables. The second score was a model-based approach similar to that of the LRS. Note that the LRS focused on predicting low mail return rates while our score focused on census undercoverage and overcoverage of housing units and persons. While we have some similar predictors in the composition of our HTE score, it is modeling a subtly different response variable.

Section 2 describes the coverage statuses that measure census undercoverage and overcoverage. Section 3 introduces the variables shown from past research to be associated with these coverage rates and how we selected those variables. Section 4 shows how we constructed the HTE scores. Section 5 compares the two HTE scores defined in Section 4 to a score based solely on not-owner-occupied status (similar to the 2010 CCM design) and the LRS. We state our conclusions and future work in Section 6.

2. Coverage Statuses

Dual system estimation is used in most post-enumeration surveys, including the 2020 PES, to produce an estimate of the true population size. This type of estimation is based on capture-recapture methodology and has been used by the Census Bureau since 1980 (Mulry and Cantwell, 2010). To measure coverage error, PES conducts an independent area-based sample of the population (known as the P sample) to compare to census housing unit and person enumerations within the same sample areas (known as the E sample). The P- and E- sample records can be placed into one of four cells shown in Figure 1.

		In P Sample (PES)		
		Yes	**No**	**Total**
In E Sample (Census)	**Yes**	N_{11}	N_{12}	N_{1+}
	No	N_{21}	N_{22}	N_{2+}
	Total	N_{+1}	N_{+2}	N

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

Assuming that the P and E samples are independent and each unit has the same chance of being in the E sample and the P sample, we can estimate our "true" population size using dual system estimation as follows:

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

Equation 1 is based on a standard Petersen (1896) or Sekar-Deming estimator to measure the size of the true population. Wolter (1986) discusses assumptions and conditions for dual system estimation. For this dual system estimate (DSE), $\widehat{N}$ is an estimator for the unknown population total N. E-sample total (N_{1+}), P-sample total (N_{+1}), and records in both samples (N_{11}) are observed.

For this paper, we calculated the DSE using E-sample correct enumeration and P-sample match coverage probabilities documented in Mule (2008) and detailed in Section 5.3. The correct enumeration probability is the probability that a person or housing unit is correctly included in the census and the match probability is the probability that a record in the P sample matches to a record in the E sample. For this research, we analyzed four coverage statuses based on the complement of these definitions:

- P-sample housing unit nonmatch
- P-sample person nonmatch
- E-sample person erroneous enumeration
- E-sample housing unit erroneous enumeration

We analyzed the association of these coverage statuses against the potential stratification variables listed in the following section to determine their predictability.

3. Variable Selection

We considered past recommendations to select which variables to test against the four coverage statuses described in the previous section. One of the sample design suggestions outlined in the 2020 Census Operational Plan (Census, 2015) included the use of the Census Bureau Planning Database for designing the 2020 PES sample. The National Research Council (2009) further recommended stratifying and oversampling in areas with large percentages of housing units or individuals in

- small multiunit structures,
- foreign-born residents,
- proxy interviews,
- whole household imputations,
- vacation homes, and
- recent additions to the Master Address File.

Such variables were recommended because housing units or individuals with these characteristics are persistently hard to count.

In the 2010 CCM, primary sampling units were stratified within each state on tenure status (i.e. owner occupied and not owner occupied), expected size (in housing units) of the primary sampling unit, and American Indians living on Reservations (for 26 states). The sample was implicitly stratified by sorting on minority status (i.e. minority and non-minority). Oversampling of some groups was needed to make sure there was enough sample to produce reliable estimates. The source of these data was the 2000 Census; thus, this information was almost 10 years old at the time of use (Moldoff, 2008).

We analyzed the association between coverage statuses and variables from four data sources:

- 2012 block-group level Planning Database (which included 5-year ACS estimates from 2006 through 2010)
- 2014 tract-level Planning Database (which included 5-year ACS estimates from 2008 through 2012)
- 2008 Master Address File
- variables based on the 2010 Census

We attempted to select only variables whose values were collected before the time of 2010 CCM; however, some of the Planning Database variables were not available for the appropriate time period. Thus, some of the Planning Database variables overlapped with the 2010 CCM time frame (i.e., the 2008-2012 ACS 5-year estimates). If we used variables from a Planning Database in practice, we would only have variable information before the sample selection of the 2020 PES primary sampling units. Using overlapping data here may overemphasize the strength of that variable if there is some deterioration over time.

Table 1 provides a list with descriptions of potential variables chosen for inclusion in the HTE scores. Multicolinearity of the variables was reduced by removing variables with high correlations to other variables researched. This list represents the variables selected after this screening was done.

Table 1: Potential Variables for Hard-to-Enumerate (HTE) Scores

Variable Name	Variable Description
Small Multi-Unit Structures	Percent of all ACS housing units within a tract that are in a structure that contains two to nine housing units.
Below Poverty Level	Percent of ACS eligible population within a tract that are classified as below the poverty level given their total family or household income within the last year, family size, and family composition.
Not High School Graduates	Percent of ACS population within a tract ages 25 years and over that are not high school graduates and have not received a diploma or the equivalent.
Not Owner Occupied	Percent of ACS occupied housing units within a tract that are not owner occupied, whether they are rented or occupied without payment of rent.
Under 5 Years Old	Percent of ACS population within a tract that is under five years old.
18-24 Years Old	Percent of ACS population within a tract that is between 18 and 24 years old.
Unemployed	Percent of ACS civilians within a tract ages 16 years and over in the labor force that are unemployed.
Foreign-Born Residents	Percent of ACS population within a tract who were not a citizen of the United States at birth.
Vacant Units	Percent of ACS housing units within a tract where no one is living regularly at the time of interview.
Crowded Housing	Percent of ACS occupied housing units within a tract with more than 1.01 persons per room.
Geocoding Errors	Percent of housing units within a tract on the Spring 2008 Master Address File that are assigned the wrong block codes.
Not on Delivery Sequence File (DSF) in Spring 2008	Percent of records within a tract not on the most recent Delivery Sequence File (i.e., did not receive United States Postal Service mail delivery at the time of the Delivery Sequence File delivery to the Census Bureau in Spring 2008).
Rural	Percent of housing units within a tract on the Spring 2008 Master Address File that were considered rural.
Proxy Responses	Percent of population within a tract from 2010 Census that were proxy respondents in nonresponse followup operations.
Complex Households	Percent of households within a tract categorized as not a "nuclear family" based on information from 2010 Census.
Whole Person Imputations	Percent of persons within a tract where information for an entire record was imputed during 2010 Census.

Each variable was summarized at the tract level and transformed into a proportion to have consistency in geography and scale. The coverage probabilities were based on the 2010 CCM data, which was collected at a block-cluster level (consisting of one or more blocks). However, the primary sampling unit for the 2020 PES will be the basic collection unit. Since the blocks and basic collection units were not delineated with the constraint of having overlapping geographical boundaries, we used tract-level information because this is the smallest geography blocks and basic collection units have in common.

We used 2010 CCM housing unit and person response data to build survey-weighted logistic regression models for each coverage status against all of the variables listed in Table 1 to determine which variables should be included in the HTE score creation. To set this up, we attached the tract-level variable information to the person and housing unit sample files that contained the coverage probabilities and survey weights. Using PROC SURVEYLOGISTIC of the SAS®[2] software, we ran these models and applied them to a list of all tracts in the U.S.

We transformed the coverage probabilities on the sample files into binary variables so they could be appropriately used as the dependent variable in the logistic regression model. For each person or housing unit k, let p_k be the probability of an E-sample correct enumeration or P-sample match with $0 \leq p_k \leq 1$ and w_k be the corresponding survey weight. Most observations had $p_k = 0$ or 1, where $p_k = 1$ represents a correct enumeration or match and $p_k = 0$ represents an erroneous enumeration or nonmatch. We split observations with non-binary probabilities into two binary observations and proportioned survey weights accordingly. For example, if a person had match probability $p = 0.8$ and a weight $w = 100$, we split the person into two records, the first with $p = 1$ and $w = 80$ and a second with $p = 0$ and $w = 20$.

All variables listed in Table 1 were merged onto each of the four datasets containing the different housing unit and person coverage probabilities and weights. Because we had tract-level variables, all persons or housing units within that tract had the same value. We evaluated the significance of the variables within each model and retained those variables that were significant for at least two of the four coverage statuses. Table 2 displays the significance of the variables from each model, with significant variables marked by 'x'. A star (*) next to a variable name indicates two or more coverage statuses with significant p-values. Seven variables that were significant for at least two of the four coverage statuses were retained for HTE score creation.

Table 2: Significant p-values for Potential Variables for Hard-to-Enumerate (HTE) Score Creation

Variable Name	Housing Unit Nonmatch	Person Nonmatch	Housing Unit Erroneous Enumeration	Person Erroneous Enumeration
Small Multi-Unit Structures			x	
Below Poverty Level				
Not High School Graduates*	x	x	x	
Not Owner Occupied*		x		x
Under 5 Years Old				
18-24 Years Old*		x		x
Unemployed				x
Foreign-Born Residents	x			
Vacant Units*	x	x	x	x
Crowded Housing				x
Geocoding Errors	x			
Not on DSF† in Spring 2008*	x	x	x	x

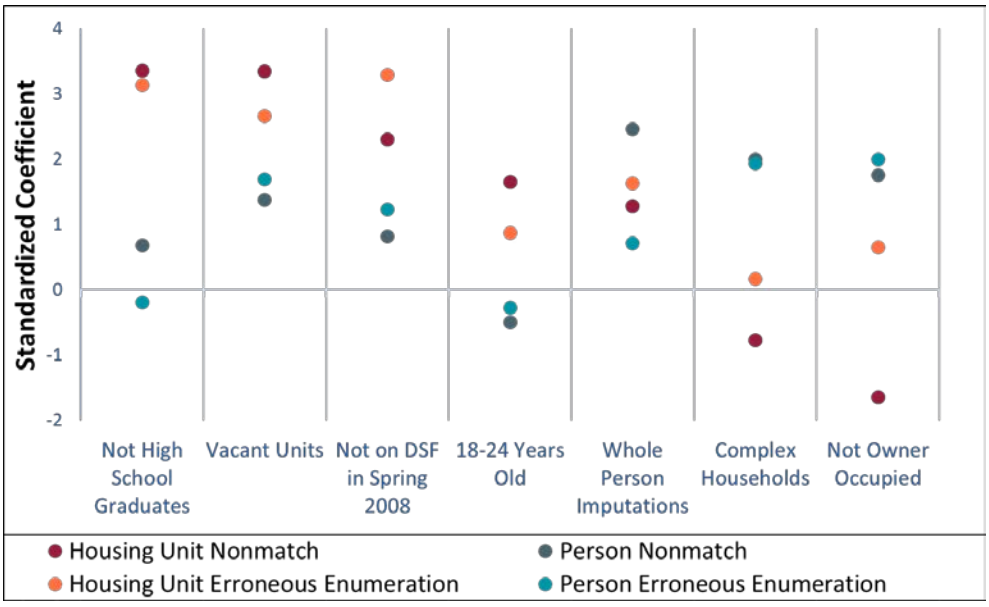
[2] Copyright © 2013 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

Rural				
Proxy Responses				x
Complex Households*		x		x
Whole Person Imputations*	x	x	x	x

†DSF: Delivery Sequence File
Data Sources: U.S. Census Bureau's 2010 Census Coverage Measurement, 2012 and 2014 Planning Databases, Spring 2008 Master Address File, and 2010 Census

To describe which variables retained had a larger positive relationship with coverage, we also produced standardized coefficients from the four models. Figure 2 shows the standardized coefficients from the four models that used only the retained variables to help visualize which variables had larger influence in the models. A larger standardized coefficient indicates a stronger positive relationship to the coverage status of interest.

We can see that the "Not High School Graduates", "Vacant Units", and "Not on DSF in Spring 2008" variables have larger influence in the models that deal with housing unit coverage. "Not Owner Occupied", "Complex Households", and "Whole Person Imputations" have larger influence in the models that focus on person coverage. Also of note is that the "Not Owner Occupied" variable is not significant in the P-sample housing unit nonmatch model and has the least influence compared to other variables on this coverage status. This variable is used as a proxy for the not-owner-occupied variable used in the 2010 CCM that was based on 2000 Census data to identify hard-to-enumerate areas. These observations motivate creating a score based on multiple variables rather that just using one variable to stratify the primary sampling units for the 2020 PES.



† DSF: Delivery Sequence File

Figure 2: Plot of Standardized Coefficients for each Coverage Status Model by Retained Variables *(Data Sources: U.S. Census Bureau's 2010 Census Coverage Measurement, 2012 and 2014 Planning Databases, Spring 2008 Master Address File, and 2010 Census)*

The following section explains how the selected variables were combined to form HTE scores.

4. Score Creation and Stratification

We created the following hard-to-enumerate scores and compared the effects of using these scores for stratification:

1) HTE Score 1 using a standardized-average approach.
2) HTE Score 2 using a model-based approach.
3) Not-owner-occupied variable.
4) LRS.

The creation and description of each score and the stratification technique is described below.

The standardized average approach was motivated by the 2001 UK version that took the sum of the variables that were in the form of proportions. Instead of simply summing the variables, here we extended the UK method by standardizing each proportion and then taking the average. Formally, for the proportion x_{it} for tract t within a selected variable i for $i = 1, \dots, n$ (with n representing the total number of selected variables), the standardized value is defined as

$$z_{it} = \frac{x_{it} - \bar{x}_i}{s_i} \tag{2}$$

where $\bar{x}_i$ and s_i are the mean and standard deviation of x_{it}, respectively. Then, the HTE Score 1 for each tract t is

$$HTE_{1,t} = \frac{\sum_{i=1}^{n} z_{it}}{n}. \tag{3}$$

These values were calculated for the entire universe of tracts.

Our second HTE score was a model-based approach similar to that of the 2011 UK hard-to-count score and the more recent LRS based on ACS data. The first step in this method was to calculate the weighted ratio of the coverage probabilities for each tract in the sample using the provided sample weights. Recall, for each person or housing unit k, $p_k = 0$ represented an erroneous enumeration or nonmatch and w_k was the corresponding survey weight. Then this weighted ratio was calculated by summing for each tract as follows:

$$\frac{\sum_{p_k=0} w_k}{\sum_{p_k=0} w_k + \sum_{p_k=1} w_k}.$$

Once these weighted ratios were calculated for each tract for the four different coverage statuses, we ran logistic regression models with the selected variables. Resulting model coefficients were applied to the entire universe of tracts, which led to four different predicted probability values for each tract. These values were then standardized and averaged in the same way as defined in equations 2 and 3 to create the HTE Score 2.

Thirdly, we used the not-owner-occupied variable from ACS 2014 tract-level Planning Database as a proxy for 2010 CCM stratification. Note that stratification using the ACS not-owner-occupied variable is different than the 2010 CCM stratification definitions. As stated previously, the 2010 CCM stratification used 2000 Census tenure information, as well as the additional block size stratification and minority status implicit stratification. However, we used the ACS not-owner-occupied variable as an approximation to make

comparisons here for simplicity. We used the HTE scores and not-owner-occupied variable of the same timeframe to reduce differences due to time lag.

Finally, we also compared these three scores to the LRS from the 2014 tract-level Planning Database. As stated in Section 1, LRS focused on predicting low mail return rates while our goal is to predict areas that are hard-to-enumerate. While not identical, we expect there to be a lot of intersection between these two scores.

To create strata using the four scores, we attempted to mimic the 2010 CCM design which placed primary sampling units that had 40 percent or more non-owner population in the harder-to-enumerate stratum (and were therefore eligible for oversampling). Applying this percentage to tracts using the not-owner-occupied variable from ACS, approximately 25 percent of the tracts would be considered harder to enumerate. We used this percentage as a threshold for dividing the tracts into high and low hard-to-enumerate strata based on the four scores. That is, for each of the four scores (HTE scores 1 and 2, LRS, and not owner occupied), we define the score-based strata variable h as

$$h = \begin{cases} high, if\ score \geq 75th\ percentile \\ low, if\ score < 75th\ percentile \end{cases}.$$

Having the same stratification threshold for each score allows us to compare the impact of stratification across scores.

As detailed in Section 5, we compared the consistency of each score using crosstabulation tables and choropleth maps. We performed an analysis of variance using PROC ANOVA of the SAS®[2] software with the newly-defined two-level score-based strata variable h to compare the effectiveness of stratification based on resulting sum of squares output. We also ran survey-weighted logistic regression models incorporating the score-based strata variable to compare significance across scores in relation to the coverage statuses. In the final analyses, we calculated standard errors of DSEs for key estimation domains using the score-based strata variable for poststratification. This analysis provided insight into how the four different stratification methods would help to reduce the variance of our coverage estimates.

5. Evaluations

With the seven retained variables, we created the two HTE scores. Recall that HTE Score 1 is the standardized-average approach and HTE Score 2 is the model-based approach. We compared these two HTE scores to the LRS and to the not-owner-occupied variable alone to answer the following questions:

1. *Do the scores agree with each other?*
2. *Does the stratification create distinct categories?*
3. *Do the score-based strata successfully capture the coverage values of interest?*
4. *Do the score-based strata reduce the variance of estimates?*

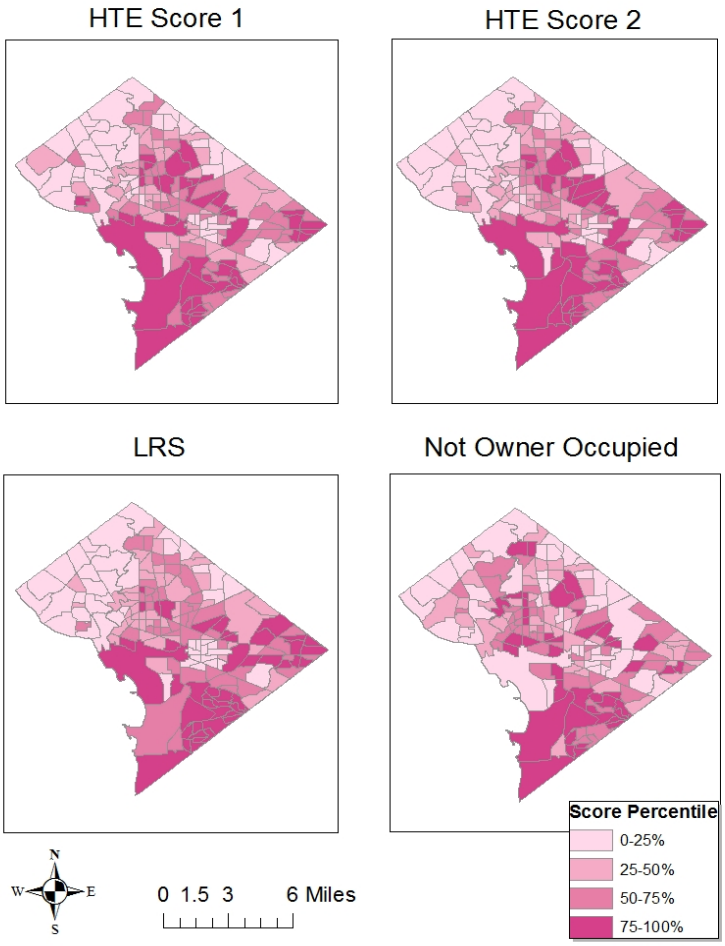
Answering these questions helped to inform the decision of whether or not using a HTE score for stratification is advantageous compared to previous methods. We answer each of these four questions in this section.

5.1 Do the scores agree with each other?

If the scores show high agreement amongst them, there is no reason to choose one above another. If the not-owner-occupied variable alone is similar to the other scores, we would

continue forward with the same stratification variable (i.e., not owner occupied) used in the 2010 CCM design. To answer this question, we explored HTE score choropleth maps and calculated the consistency within HTE strata across the four scores.

First, we looked at some example states' maps to visually compare the scores of each tract within a state. Figure 3a shows four different scores for the District of Columbia. Darker pink indicates a higher score. The darkest pink in each map refers to tracts in the top 25 hard-to-enumerate percentile of the corresponding score. From Figure 3a, the HTE scores 1 and 2 are very closely aligned. The tracts with higher scores from LRS are more clustered to southern portion of the map than HTE scores 1 and 2, while the not owner occupied seems to be more scattered. While there are differences, we also see consistency among all four maps.

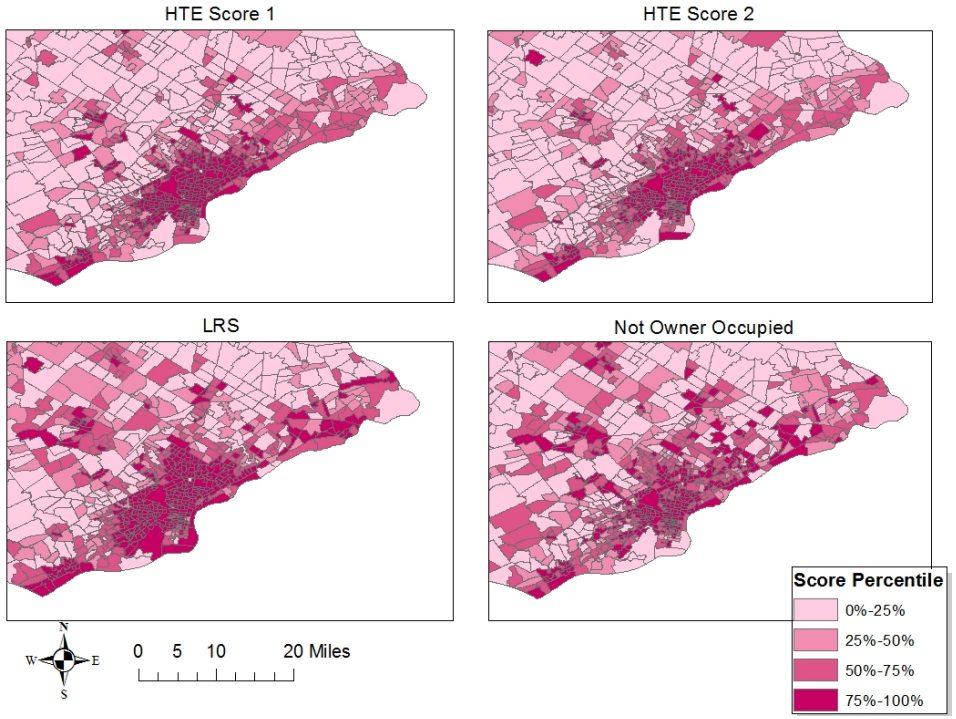


†HTE: Hard-to-Enumerate, LRS: Low Response Score

Figure 3a: Choropleth Maps of Scores by Tract for the District of Columbia *(Data Sources: U.S. Census Bureau's 2010 Census Coverage Measurement, 2012 and 2014 Planning Databases, Spring 2008 Master Address File, 2010 Census, and 2010 TIGER/Line Shapefiles)*

Since the District of Columbia is relatively compact and urban, we also looked at other locations that had a mixture of urban, suburban, and rural areas. Figures 3b and 3c give

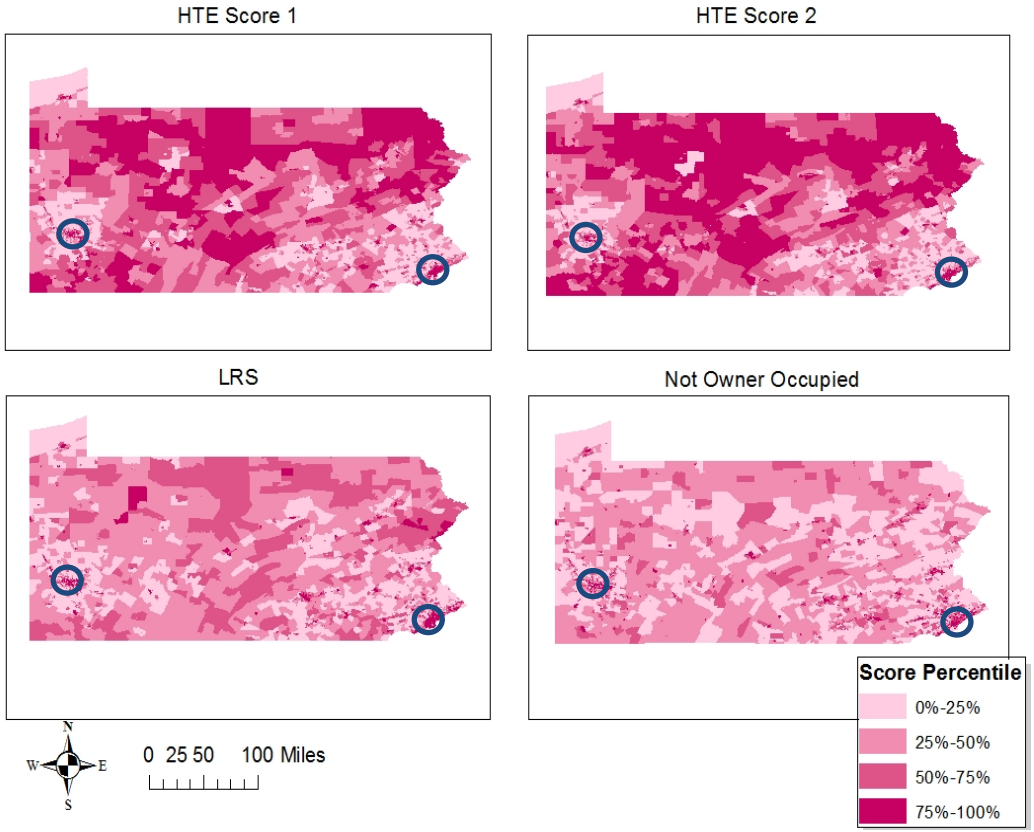
another example comparing the four scores for tracts in Pennsylvania. For all four maps, there are clusters of high hard-to-enumerate tracts in the city centers (e.g., Philadelphia in Figure 3b).



†HTE: Hard-to-Enumerate, LRS: Low Response Score

Figure 3b: Choropleth Maps of Scores by Tract for Pennsylvania (Close-up of Philadelphia) (*Data Sources: U.S. Census Bureau's 2010 Census Coverage Measurement, 2012 and 2014 Planning Databases, Spring 2008 Master Address File, 2010 Census, and 2010 TIGER/Line Shapefiles)*

In Figure 3c, tract boundaries were removed so that small tracts were not obscured. Again, darker pink indicates a higher score. Note that HTE scores 1 and 2 look like they have more dark pink, but this is not the case; these scores are just identifying hard-to-enumerate tracts that happen to have a larger land area. Figure 3c shows HTE scores 1 and 2 identifying areas in the North as being in the upper percentiles of hard-to-enumerate. These areas are more rural countryside. The LRS and not owner occupied are not identifying these areas as being in the upper 25th percentile. Depending on the threshold chosen for stratification, these tracts may not be included in the hard-to-enumerate stratum and therefore eligible for oversampling. We see similar trends in other states, with all four scores consistently identifying similar hard-to enumerate areas in urban areas; however, the HTE scores 1 and 2 also identify hard-to-enumerate areas in more rural areas.



†HTE: Hard-to-Enumerate, LRS: Low Response Score

Figure 3c: Choropleth Maps of Scores by Tract for Pennsylvania (*Data Sources: U.S. Census Bureau's 2010 Census Coverage Measurement, 2012 and 2014 Planning Databases, Spring 2008 Master Address File, 2010 Census, and 2010 TIGER/Line Shapefiles)*

This is further reflected in Figure 3d, which gives the percentiles of tracts containing rural areas in Pennsylvania using the percent rural variable from the 2014 Planning Database. We created strata for percent rural similar to how we created strata for the HTE scores, with the top 25 percent rural tracts in the high hard-to-enumerate stratum and the lower 75 percent rural tracts in the low hard-to-enumerate stratum. Many of the dark pink areas, representing tracts containing high percent of rural area, match up to the dark pink areas in Figure 3c that have high scores for HTE scores 1 and 2.

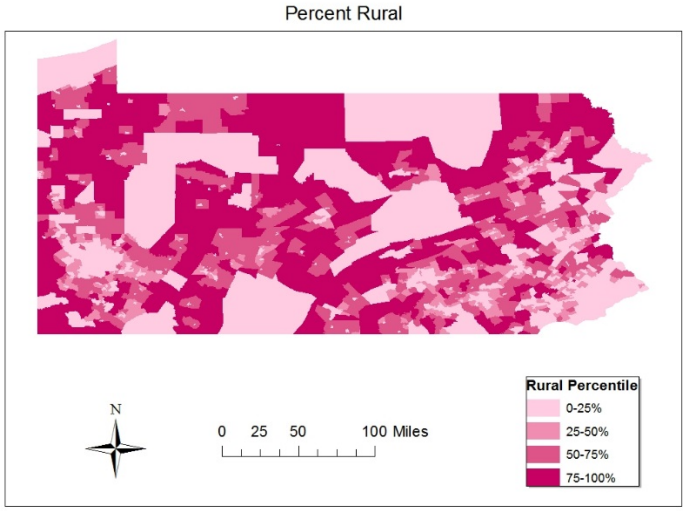


Figure 3d: Choropleth Maps of Percent Rural by Tract for Pennsylvania (*Data Sources: U.S. Census Bureau's 2014 Planning Database and 2010 TIGER/Line Shapefiles)*

To quantify what we saw in the maps, we calculated the agreement between the scores based on whether or not a tract is placed in the high or low strata for all of the tracts in the nation. Recall from Section 3 that the top 25 percent of tracts are placed in the high hard-to-enumerate stratum for each of the four scores. Table 3 shows that the score-based strata using HTE scores 1 and 2 are very similar, with 88.4 percent agreement of which tracts are within the high hard-to-enumerate stratum and 94.2 percent overall agreement. HTE Score 1 agrees with not-owner-occupied and LRS strata at a higher percent than HTE Score 2. The strata for both HTE scores 1 and 2 agree with the LRS and not-owner-occupied strata 77.9 percent to 87.4 percent of the time overall. However, within the high hard-to-enumerate stratum alone, we see lower agreement, with 55.8 percent to 74.7 percent of the time. Overall, all scores are consistent with each other more often than not (i.e., more than 50 percent of the time).

We also looked at the percent agreement between the four scores and the percent rural variable to quantify what we observed in Figure 3c. We created strata for percent rural similar to how we created strata for the scores, with the top 25 percent rural tracts in the high hard-to-enumerate stratum and the lower 75 percent rural tracts in the low hard-to enumerate stratum. While the agreement scores are much lower overall, the HTE scores 1 and 2 both agree at a higher rate with percent rural than not owner occupied and LRS. This indicates that the HTE scores 1 and 2 are capturing more tracts with a high percent of rural compared to the other two scores.

Table 3: Percent Agreement between Two Scores within HTE† Strata

Comparison	Overall Percent Agreement (within Low or High Stratum)	Percent Agreement within High Stratum
HTE Score 1 vs HTE Score 2	94.2	88.4
HTE Score 1 vs Not Owner Occupied	82.2	64.3
HTE Score 2 vs Not Owner Occupied	77.9	55.8
LRS† vs Not Owner Occupied	86.4	68.2
HTE Score 1 vs LRS	87.4	74.7
HTE Score 2 vs LRS	83.6	67.2
Percent Rural vs HTE Score 1	59.5	19.0
Percent Rural vs HTE Score 2	62.3	24.6
Percent Rural vs LRS	53.1	6.2
Percent Rural vs Not Owner Occupied	51.7	3.3

†HTE: Hard-to-Enumerate, LRS: Low Response Score
Data Sources: U.S. Census Bureau's 2010 Census Coverage Measurement, 2012 and 2014 Planning Databases, Spring 2008 Master Address File, and 2010 Census

5.2 Does the stratification create distinct categories?

We checked that each score-based stratification creates two categories that are significantly different from each other. If this is not the case, we would not want to use that stratification. We performed an analysis of variance using PROC ANOVA of the SAS®[2] software with the 2010 CCM housing unit and person sample data and the two-level score-based strata variable h of high hard-to-enumerate ($\geq 75th\ percentile$) and low hard-to-enumerate ($< 75th\ percentile$). This was done to verify that the split divides tracts into two significantly different categories for the four coverage statuses. We also compared how well this split performs across the four scores.

Table 4 gives some resulting ANOVA output that decomposes the total variation into the between-group variance and within-group variance. Larger mean squared error between groups (MSB) and larger F statistics indicate more differences in the scores of tracts *between* the different strata. A smaller mean squared error within groups (MSW) indicates the tract scores are similar *within* a stratum. So, greater gain in precision arise from stratification that yields large MSB and small MSW. The results show that the split is significant across all scores for each coverage status, with p-values much less than 0.1.

The results are mixed for the score comparison across coverage statuses. For P-sample housing unit nonmatch and E-sample housing unit erroneous enumeration coverage statuses, strata based on HTE scores 1 and 2 have larger MSB and F statistics compared to strata based on LRS and not owner occupied. However, for the E-sample person erroneous enumeration, the not-owner-occupied and LRS scores have larger MSB and F statistics compared to the HTE scores 1 and 2. MSW remains about the same across scores. Nevertheless, the split at the 75th percentile seems to adequately divide tracts into high and low hard-to-enumerate strata for all four scores and coverage statuses.

Table 4: Analysis of Variance (ANOVA) Statistics for Score-Based Strata (Unweighted)

Coverage Status	Score	MSW†	MSB†	F Statistic	P-Value
P-Sample Housing Unit Nonmatch	HTE† Score 1	0.04	35.4	990.8	<.0001
	HTE Score 2	0.04	42.1	1,181.4	<.0001
	Not Owner Occupied	0.03	0.4	10.9	0.0009
	LRS†	0.04	10.3	285.9	<.0001
P-Sample Person Nonmatch	HTE Score 1	0.10	272.6	2,619.6	<.0001
	HTE Score 2	0.10	250.2	2,403.4	<.0001
	Not Owner Occupied	0.10	168.6	1,616.0	<.0001
	LRS	0.10	273.8	2,631.8	<.0001
E-Sample Housing Unit Erroneous Enumeration	HTE Score 1	0.03	34.9	1,007.7	<.0001
	HTE Score 2	0.03	46.8	1,352.2	<.0001
	Not Owner Occupied	0.03	0.4	12.3	0.0005
	LRS	0.03	12.6	363.0	<.0001
E-Sample Person Erroneous Enumeration	HTE Score 1	0.11	233.4	2,185.6	<.0001
	HTE Score 2	0.11	193.8	1,812.8	<.0001
	Not Owner Occupied	0.11	281.2	2,636.1	<.0001
	LRS	0.10	289.0	2,709.8	<.0001

†HTE: Hard-to-Enumerate, LRS: Low Response Score, MSW: Mean Squared Error within groups, MSB: Mean Squared Error between groups

Data Sources: U.S. Census Bureau's 2010 Census Coverage Measurement, 2012 and 2014 Planning Databases, Spring 2008 Master Address File, and 2010 Census

5.3 Do the score-based strata successfully capture the coverage values of interest?

We answered this question in two different ways. First, we ran a survey-weighted logistic regression model to compare the relationship between each of the four scores with their defined strata to the four coverage statuses. Second, we calculated poststratified estimates of housing unit match and correct enumeration rates to see which scores captured the harder-to-enumerate tracts in their stratification.

For each score, the survey-weighted logistic regression model used a score-based strata variable as the specified model stratification and the corresponding raw tract scores as the only independent variable in the model. We performed this analysis for each coverage status as the dependent variable for a total of 16 survey-weighted logistic regression models.

The HTE scores 1 and 2 give percent concordance higher than the LRS and not owner occupied for P-sample housing unit nonmatch and E-sample housing unit erroneous enumeration. This indicates a stronger relationship between HTE scores 1 and 2 with housing unit coverage compared to the other two scores. We see more agreement among scores for person coverage.

Table 5: Survey-Weighted Logistic Regression Models Output with Tract Scores as the Independent Variable

Coverage Status	Score	Estimate	Std. Error	*t* Statistic	P-Value	Percent Concor-dance
P-Sample Housing Unit Nonmatch	HTE† Score 1	0.70	0.09	8.1	<.0001	60.3
	HTE Score 2	0.35	0.04	8.3	<.0001	60.3
	Not Owner Occ.	<0.00	<0.00	-0.1	0.9426	26.5
	LRS†	0.03	0.01	3.6	0.0003	50.3
P-Sample Person Nonmatch	HTE Score 1	0.72	0.03	24.7	<.0001	58.5
	HTE Score 2	0.38	0.02	21.5	<.0001	57.7
	Not Owner Occ.	0.01	<0.00	8.7	<.0001	48.8
	LRS	0.06	<0.00	19.7	<.0001	58.0
E-Sample Housing Unit Erroneous Enumeration	HTE Score 1	0.91	0.07	12.4	<.0001	61.8
	HTE Score 2	0.39	0.05	8.4	<.0001	61.5
	Not Owner Occ.	0.01	<0.00	2.9	0.0043	42.4
	LRS	0.05	0.01	6.1	<.0001	54.1
E-Sample Person Erroneous Enumeration	HTE Score 1	0.58	0.03	20.3	<.0001	56.9
	HTE Score 2	0.31	0.02	16.5	<.0001	55.8
	Not Owner Occ.	0.01	<0.00	12.3	<.0001	52.5
	LRS	0.04	<0.00	17.0	<.0001	56.8

†HTE: Hard-to-Enumerate, LRS: Low Response Score
Data Sources: U.S. Census Bureau's 2010 Census Coverage Measurement, 2012 and 2014 Planning Databases, Spring 2008 Master Address File, and 2010 Census

We also compared the four scores by calculating the poststratified match and correct enumeration rates that are used to create the housing unit DSE (detailed in Section 2). In the poststratification process, we calculated the weighted number of P-sample housing unit matches and the weighted number of E-sample housing unit correct enumerations for each tract. We calculated these by multiplying the housing unit match and correct enumeration probabilities by their corresponding weights and summing them up to the tract level (i.e., for each tract, we calculated $\sum_{k=1}^{n} p_k w_k$ for housing units $k = 1, \ldots, n$ with p_k and w_k representing the housing unit probability and weight, respectively). Then we summed these values to the poststratum level (i.e., within high hard-to-enumerate stratum and low hard-to-enumerate stratum). Using these pieces, we calculated the housing unit DSE within each stratum as

$$DSE_h = Census_h * \frac{\sum_t ce_{ht} / \sum_t we_{ht}}{\sum_t match_{ht} / \sum_t wp_{ht}}, \quad (4)$$

where

- t represents the tracts within stratum h,
- $Census_h$ represents the 2010 Census housing unit counts within stratum h,
- $match_{ht}$ represents the weighted number of P-sample housing unit matches for each tract t within stratum h,
- ce_{ht} represents the weighted number of E-sample housing unit correct enumerations for each tract t within stratum h,
- we_{ht} represents the total E-sample weights for each tract t within stratum h, and

- wp_{ht} represents the total P-sample weights for each tract t within stratum h.

We can sum across strata to get a total DSE. Table 6 shows the match rate and correct enumeration rate within the high hard-to-enumerate stratum and low hard-to-enumerate stratum, where the match rate and correction enumeration rate is defined by the denominator and numerator, respectively, from Equation 4. We statistically compared these estimates using HTE Score 1, HTE Score 2 and LRS to not owner occupied (recall we are considering not owner occupied as our proxy for the characteristic used to stratify the primary sampling units in the 2010 CCM.) A (*) in the table indicates that the value is significantly different from the not-owner-occupied value at the 0.1 level.

The match and correct enumeration rates for the HTE scores 1 and 2 and the LRS are significantly lower than the not-owner-occupied match and correct enumeration rates within the high hard-to-enumerate strata. This means that the high strata that we created using the HTE scores 1 and 2 are capturing tracts that have lower match and correct enumeration rates (i.e., the tracts that are harder to enumerate) better than the high stratum for the not owner occupied. Conversely, we also see significantly higher match and correct enumeration rates for HTE scores 1 and 2 and the LRS compared to not owner occupied in the low hard-to-enumerate strata (except for the correct enumeration rate using LRS). This further motivates the idea that we are capturing the harder-to-enumerate tracts using the HTE scores 1 and 2 better than using not owner occupied alone.

Table 6: Housing Unit Match and Correct Enumeration Rates Using Score-Based Strata Variable for Poststratification

Statistic	HTE† Score 1	HTE Score 2	LRS†	Not Owner Occupied (Baseline)
High HTE Stratum Match Rate	95.5*	95.4*	96.5*	97.0
Low HTE Stratum Match Rate	97.5*	97.6*	97.2*	97.0
High HTE Stratum Correct Enumeration Rate	95.4*	95.2*	96.4*	96.9
Low HTE Stratum Correct Enumeration Rate	97.9*	98.0*	97.5	97.4

†HTE: Hard-to-Enumerate, LRS: Low Response Score
Data Sources: U.S. Census Bureau's 2010 Census Coverage Measurement, 2012 and 2014 Planning Databases, Spring 2008 Master Address File, and 2010 Census

5.4 Do the score-based strata reduce the variance of estimates?

We would like to see some variance reduction of key domain estimates using HTE scores 1 and 2 or LRS compared to the not-owner-occupied score. Table 7 displays the standard error of DSEs for various key domains of interest, once again using the poststratification from the four scores described in Equation 4 of Section 5.3. These standard errors were calculated using successive difference replication variance estimation, derived in Fay and Train (1995). In this table, we provide the estimated standard errors for domains such as housing unit status, housing unit type, and bilingual status.

We would like to see some variance reduction of hard-to-enumerate groupings within these key domains, such as renters, vacant units, small multi-units, or bilingual blocks with our HTE scores 1 and 2. Once again, a (*) in the table indicates that the value is significantly different from the not-owner-occupied value at the 0.1 level. However, Table 7 shows that

the standard errors for HTE scores 1 and 2 and the LRS are not significantly different from the not-owner-occupied score for any of these domains.

Table 7: Standard Errors of Housing Unit Dual System Estimates (DSEs) for Key Domains

Domain	DSE Poststrata	HTE† Score 1	HTE Score 2	LRS†	Not Owner Occupied (Baseline)
Tenure	Owner	83	82	82	83
	Renter	129	129	129	128
	Vacant	138	138	134	134
Housing Unit Type	Single Unit	111	110	110	110
	Small Multi-Unit (2-9)	75	76	75	74
	Large Multi-Unit (10+)	143	143	144	145
	Trailer/Other	75	75	77	75
Bilingual Status	Not Bilingual Block	198	198	197	195
	Bilingual Block	58	59	58	57

†HTE: Hard-to-Enumerate, LRS: Low Response Score
Data Sources: U.S. Census Bureau's 2010 Census Coverage Measurement, 2012 and 2014 Planning Databases, Spring 2008 Master Address File, and 2010 Census

6. Conclusion

In conclusion, we looked at potential stratification variables from several data sources and compared their association with four coverage statuses in order to see if improvements could be made to the stratification of the 2020 PES sample design. From a select group of these variables, we created two versions of the HTE score and compared them against not-owner-occupied and LRS scores.

Overall, the two HTE scores yield similar results to each other and could be used interchangeably. An advantage of HTE Score 1 is that it is straightforward and easy to explain. However, since this score is an average of selected variables, it is sensitive to any additional variables. Variables would first have to be screened for high association with coverage; otherwise, the variable could be detrimental to the score. On the other hand, adding variables to HTE Score 2 will not harm this score since it uses these variables in a model. The HTE Score 2 also more closely aligns to what was done more recently in the UK and the most recent LRS.

Our observations indicated that the HTE scores, not owner occupied and LRS give similar results in more urban areas, while the HTE scores are capturing more rural areas than the other two scores in the high hard-to-enumerate strata. We also better capture tracts with lower match and correct enumeration rates in the high hard-to-enumerate strata using our HTE scores 1 and 2. However, evaluations based on survey-weighted logistic regression models, analysis of variance, and standard errors of housing unit DSEs for key domain estimates all indicate that the HTE scores give comparable results to the not-owner-

occupied and LRS scores in terms of how well we are predicting census coverage for the country as a whole. The decision was made based on preliminary information to continue to use tenure status (i.e., owner occupied and not owner occupied) as the stratification variable for 2020 PES similar to how it was used in 2010 CCM.

There were some limitations to this research which, if addressed, could help inform future work. As stated in Section 3, this research was done using tract-level information since this is the smallest geography basic collection units from 2020 Census and collection blocks from 2010 Census have in common. At the time of this research, the relationship file between basic collection units and blocks had not yet been developed. A more detailed HTE score at a lower geography may improve upon the scores presented here. In the future we may like to focus on other potential stratification variables, including the type of enumeration area, metropolitan statistical area, 2020 Census address canvassing status, or available variables from administrative records.

Acknowledgements

Thank you to Laura Davis and Vincent T. Mule for their comments and input on this document.

References

Bruce, A., Robinson, G., and Sanders, M. (2001), "Hard-to-Count Scores and Broad Demographic Groups Associated with Patterns of Response Rates in Census 2000". In Proceedings of the Social Statistics Section, American Statistical Association.

Erdman, C. and Bates, N. (2017). "The Low Response Score (LRS), A Metric to Locate, Predict, and Manage Hard-to-Survey Populations." Public Opinion Quarterly, Vol. 81, No. 1, pp. 144–156.

Fay, R. and Train, G. (1995). "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties," Proceedings of the Government Statistics Section, Alexandria, VA: American Statistical Association, pp.154-159.

Moldoff, M. (2008). "The Design of the Coverage Measurement Program for the 2010 Census." DSSD 2010 Census Coverage Measurement Memorandum Series #2010-B-07, Retrieved from https://www.census.gov/coverage_measurement/pdfs/2010-B-07.pdf.

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.

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.

Office for National Statistics (2000), "One Number Census Methodology" ONC(SC)00/11, June 2000.

Office for National Statistics (2009), “Predicting patterns of household non-response in the 2011 Census”. Census Advisory Group paper AG(09)17.

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.

Schellhamer, T. (2010), “2010 Census Post-Enumeration Survey: Differential Sampling Research and Methodology Results,” DSSD 2010 Census Coverage Measurement Memorandum Series #2010-C-22, April 1, 2010.

U.S. Census Bureau (2015), “2020 Census Operational Plan,” Version 1.1, November 2015.

Wolter, Kirk M. (1986), “Some Coverage Error Models for Census Data”. Journal of the American Statistical Association, Vol. 81, No. 394, pp. 338- 346.