Relationship between positive responses to child-specific probes on 2010 Census questionnaires and 2010 Census Coverage Measurement nonmatching young children

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

Abstract
As part of the preparations for the 2020 U.S. Census, the U.S. Census Bureau conducted research to gain insight about the causes of the long-standing undercount of young children in decennial censuses and in Census Bureau sample surveys. This paper reports results of regression modeling to explore the relationship between the positive responses to child-specific probes on the 2010 Census questionnaire and 2010 Census Coverage Measurement (CCM) weighted nonmatching children ages 0 to 4. The units of analysis are Tapestry segments from a third-party segmentation of the population by geography and lifestyle. The coverage probes for young children and other household members appeared for the first time on the 2010 Census questionnaire so analyses of comparable data from previous censuses are not available. The paper also examines whether most positive responses to coverage probes are in areas less likely to respond to the census or if some respondents in areas with high response are expressing confusion about whether to list children on their census forms.

Key words: 2010 U.S. Census, undercount, Low Response Score, Tapestry segmentation, coverage probes

1. Introduction

The first evaluation of the coverage of a U.S. Census detected an undercount of young children in the 1950 U.S. Census using demographic methods (Coale 1955). Next, the U.S. Census Bureau evaluated the 1960 U.S. Census with a new methodology, called Demographic Analysis, that uses vital records to form an estimate of the total population constructed independently from the census in that it did not use current or past census counts (Seigel and Zelnick 1966). The U.S. Census Bureau has continued to improve and refine the Demographic Analysis methodology and use it as an evaluation tool for evaluating coverage error in censuses, including the 2010 U.S. Census (U.S. Census Bureau 2010, 2012). One advantage of the methodology is that it is able to produce estimates of census coverage for single years of age since the vital records used in forming estimates are available nationwide. West and Robinson (1999) identified a persistent pattern of a net undercount of young children, those ages 0 to 4, for U.S. Censuses from 1950 to 1990 based on the Demographic Analysis estimates for census coverage error.

[1] This report is released to inform interested parties and encourage discussion of work in progress. The views expressed on statistical, methodological, and operational issues are those of the author and not necessarily those of the U.S. Census Bureau.

Prior to the 2010 U.S Census, the Demographic Analysis estimates of the undercount of young children had not been corroborated by the U.S. Census Bureau's other method of measuring census coverage error, the Post Enumeration Survey (PES). The PES methodology involves performing an independent enumeration of a sample of blocks that is matched to the census enumerations. Then the program uses the results of the matching operation in the dual system estimator to form estimates of population size for estimating census coverage. Since the estimates are based on a sample, the number of demographic categories used in forming census coverage estimates had to be limited. Estimates were made for age groups, not single years of age, which meant that separate estimates of coverage were not available for young children for individual years of age. However, for the 2010 Census, the U.S. Census Bureau conducted a PES as part of the 2010 Census Coverage Measurement Program (CCM) and changed the dual system estimation methodology from one based on poststratification to one based on logistic regression, which enabled forming estimates for more groups than previously possible. The 2010 CCM results indicated a very small net overcount of 0.01 percent for the nation as whole. Net undercounts were observed for adult males in two categories, ages 18-29 and ages 25-34, and children 0-4 years of age, but a net overcount for all other age-sex categories although the estimated overcounts for children ages 5 -10 years and females ages 18-29 years were not statistically significant from zero (Mule 2012).

The undercount of young adult males had been observed in previous U.S. Census coverage estimates from both Demographic Analysis and PES methodologies. Young adult males tend to be highly mobile which makes them difficult to count. However, young children are dependent on adults to care for them so there were no obvious reasons for young children to be difficult to count as discussed in O'Hare (2015).

The new evidence from the 2010 CCM increased concerns about the undercount of young children and led the U.S. Census Bureau to form the Research Team on the Undercount of Young Children to investigate further. The Research Team focused on examining the characteristics of children ages 0 to 4 who were missed in the 2010 Census with the goal of gathering information that would help to improve the 2020 U.S. Census count of young children.

The Research Team conducted analyses with a wide range of census and survey data and a commercial marketing database to look for patterns that would identify sociodemographic and geographic characteristics of young children missed by the census and aid in explaining why these young children are not counted (Griffin and Konicki 2017b, 2017c). A summary of the research by the team members with references appears in Konicki and Griffin (2018). As a result of this research, steps are being taken to improve the count of young children in the 2020 Census (Jarmin 2018).

The analyses in this paper leverage the Research Team's findings and a third-party commercial marketing database. The analyses explore whether areas where Census respondents indicate uncertainty about whether to include one or more young children on their census form also are likely to display Census coverage problems for young children. In addition, the paper explores whether areas that tend to have low survey and census response rates are also the areas that have problems with the enumeration of young children.

2. Data

Our analyses use data from the 2010 Census questionnaire, the 2010 Census Coverage Measurement Program, and the Low Response Score (LRS) that was developed by the U.S. Census Bureau, and a third-party classification into ‘segments’ reflecting lifestyle.

2.1 Esri Tapestry segmentation

One of the data sources for our analyses is an external third-party database offered by Esri, a company that has worked with the U.S. Census Bureau for years to provide mapping and spatial analytics software. The Esri product we use in our analyses is a geographic and lifestyle segmentation known as Tapestry^{TM}. Tapestry has a long history of use in marketing goods and services and is not commonly used in survey research or census taking. However, recent research has shown that it may be useful in planning the communications campaign for the 2020 Census and may improve survey response propensity models (Mulry et al. 2018).

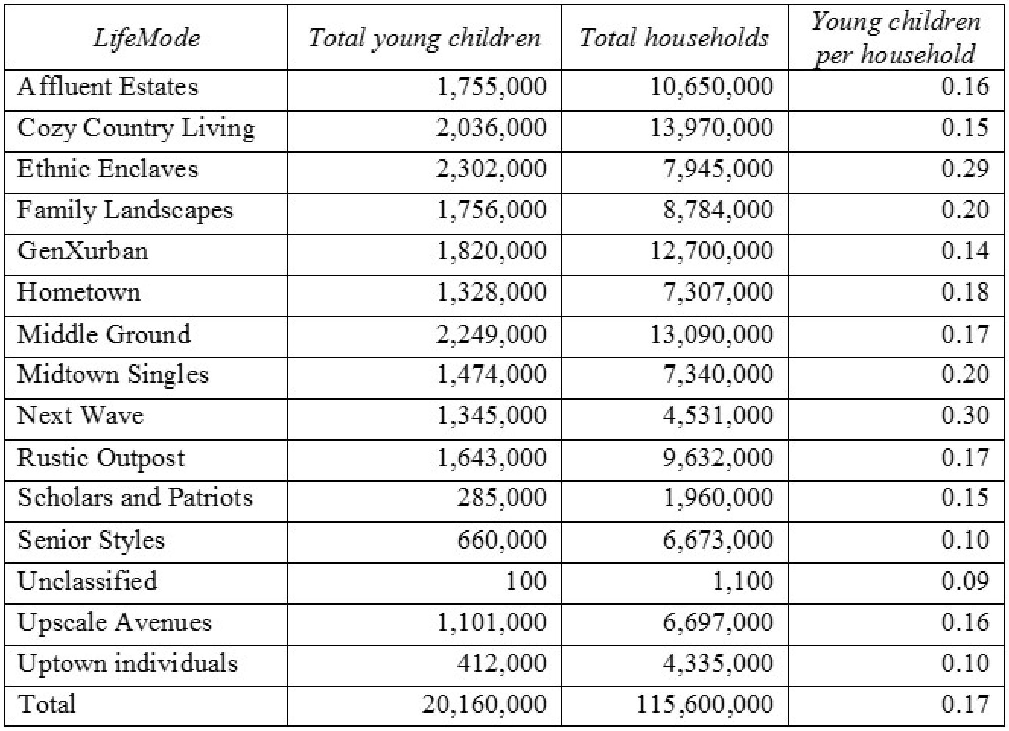
Although Tapestry is a proprietary product, the documentation does give the data sources as including the 2010 Census, the American Community Survey(ACS), Experian’s ConsumerView database (2018) and the Survey of the American Consumer from GfK MRI (Esri 2017). The methodology for forming the segments uses cluster analysis and data mining, and an updated version is produced annually.

For each of the 67 segments, Tapestry provides information useful in designing communications tailored for the residents. Included are attributes that distinguish the segment, such as a score on a race and diversity index, age by sex distribution, income and net worth, and neighborhood and socioeconomic traits. Other data includes household compositions, vacancy rates, housing structures, employment indices, government program participation rates, financial well-being, media consumption habits and internet usage. The segments receive names that reflect the lifestyle of the residents, such as Dorms to Diplomas, Small Town Simplicity, and Golden Years.

The segments aggregate to 14 LifeModes that are constructed to reflect lifestyle and life stage by grouping people who share a common experience, such as born in the same generation, or immigration from another country, or a significant socioeconomic trait, like affluence. The segments also aggregate to six urbanization groups that reflect area features such as population density and size of city. Detailed descriptions of the segments and their assignments to LifeModes and urbanization groups may be found on the Esri website (Esri 2017). To illustrate, Table 1 contains the distribution of young children counted in the 2010 Census by LifeMode. In this paper, we define *household* as all the people living in an occupied housing unit. As Table 1 shows, we use the nationwide total of 115.6 million households in our analyses. The total excludes addresses which received questionnaires that did not have the coverage probes as part of census experiments and addresses where the census population count was imputed (Griffin and Konicki 2017a).

Geographic assignments of segments depend on the characteristics of individuals and are available for several types of geographies, such as Census block groups[2], Census tracts[3] and zip codes. Initially, Census block groups are assigned to segments. Then each Census tract is assigned to the segment that is dominant among the block groups within it. In our analyses, we use the segments assigned at the Census tract level. Then housing units are assigned to the segment of their tract. Consequently, all the housing units in a tract are in the same segment.

Table 1. Young Children in 2010 Census by Tapestry LifeMode



LifeMode	*Total young children*	*Total households*	*Young children per household*
Affluent Estates	1,755,000	10,650,000	0.16
Cozy Country Living	2,036,000	13,970,000	0.15
Ethnic Enclaves	2,302,000	7,945,000	0.29
Family Landscapes	1,756,000	8,784,000	0.20
GenXurban	1,820,000	12,700,000	0.14
Hometown	1,328,000	7,307,000	0.18
Middle Ground	2,249,000	13,090,000	0.17
Midtown Singles	1,474,000	7,340,000	0.20
Next Wave	1,345,000	4,531,000	0.30
Rustic Outpost	1,643,000	9,632,000	0.17
Scholars and Patriots	285,000	1,960,000	0.15
Senior Styles	660,000	6,673,000	0.10
Unclassified	100	1,100	0.09
Upscale Avenues	1,101,000	6,697,000	0.16
Uptown individuals	412,000	4,335,000	0.10
Total	20,160,000	115,600,000	0.17

Source: Griffin and Konicki (2017b) Table 4. Rounded for disclosure avoidance.

2.2 Positive responses to 2010 Census coverage probes

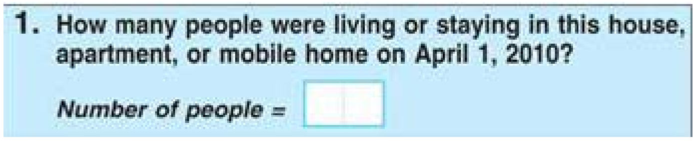
The coverage probes for young children and other household members appeared for the first time on the 2010 Census questionnaire and therefore, comparable data are not available from previous censuses. Positive responses to the coverage probes provide a new source of data for studying enumeration problems, particularly for young children. Both the self-response and enumerator-administered questionnaires included these probes, which identified addresses where the respondent acknowledged not including one or more people living or staying at the address on the census questionnaire.

Figures 1 and 2 show how the coverage probes appeared on the self-response questionnaire. Mail was the primary self-response mode in 2010. Telephone self-responses also were

[2] Census block groups are statistical divisions of census tracts, are generally defined to contain between 600 and 3,000 people, and are used to present data and control block numbering. A block group consists of clusters of blocks within the same census tract (U.S. Census Bureau, 2017).

[3] Census tracts are small areas meant to represent neighborhoods generally with a population between 1,200 and 8,000 people with an optimum size of 4,000 people and formed in collaboration with local officials (U.S. Census Bureau, 2017).

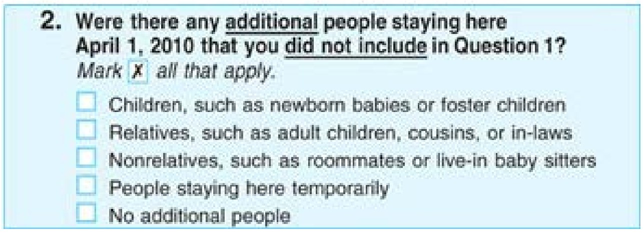
accepted at a call center but few people took advantage of this option. The first question, shown in Figure 1, asked for the number of people living or staying at the residence. The following question, shown in Figure 2, contained the coverage probes that asked if there were additional people not included in the count given in the first question. There are four categories of people who might not have been included in the count in addition to the response "no additional people." The positive responses to the category "children, such as newborn babies or foster children" is the one we use in our analyses. After the questions shown in Figures 1 and 2, the questions that collected names and characteristics of the people at the address were asked.



1. How many people were living or staying in this house, apartment, or mobile home on April 1, 2010?

Number of people = □□

Figure 1. Facsimile of population count question
on the self-response questionnaire for the 2010 Census

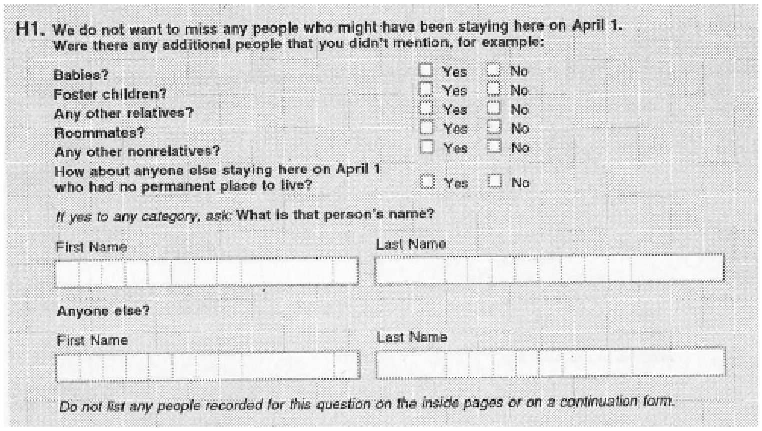


2. Were there any additional people staying here April 1, 2010 that you did not include in Question 1?
Mark ☒ *all that apply.*

□ Children, such as newborn babies or foster children
□ Relatives, such as adult children, cousins, or in-laws
□ Nonrelatives, such as roommates or live-in baby sitters
□ People staying here temporarily
□ No additional people

Figure 2. Facsimile of undercount question
on the self-response questionnaire for the 2010 Census

The coverage probes were asked in the enumerator-administered questionnaires during the 2010 Census Nonresponse Followup at addresses that did not self-respond and at addresses in areas where mail delivery was not practical so Census workers conducted interviews, called Update/Enumerate. However, the coverage probes appeared at a different point in the questionnaire. The question asking for the count of the number of people at the address appeared first, and then the questions concerning the names and characteristics of the people living at the address. Next, the enumerator asked the coverage probes shown in Figure 3.



**H1. We do not want to miss any people who might have been staying here on April 1.
Were there any additional people that you didn't mention, for example:**

Babies?	□ Yes	□ No
Foster children?	□ Yes	□ No
Any other relatives?	□ Yes	□ No
Roommates?	□ Yes	□ No
Any other nonrelatives?	□ Yes	□ No
How about anyone else staying here on April 1 who had no permanent place to live?	□ Yes	□ No

If yes to any category, ask: **What is that person's name?**

First Name

Last Name

Anyone else?

First Name

Last Name

Do not list any people recorded for this question on the inside pages or on a continuation form.

Figure 3. 2010 Census Nonresponse Followup and Update/Leave
enumerator-administered questionnaire

The 2010 Coverage Followup (CFU) attempted to contact the households that gave positive responses to the coverage probes for a telephone interview to determine whether the persons should be added to the census. Across the U.S., 611,606 households gave a positive response to the child-specific probe, and CFU was able to add 69,383 young children (Griffin and Konicki 2017b). However, CFU did not attempt to interview households that did not provide a telephone number on their census questionnaire, and CFU was unable to contact some households despite having a telephone number for them. Therefore, CFU was unable to determine whether many of the households that gave a positive response to the child-specific probe contained children that should have been counted in the census with the household.

For this study, we focus on the positive responses to the child-specific coverage probes instead of additions to avoid possible confounding that might be caused by nonresponse in the Coverage Followup. After all, additions require that a Coverage Followup interview occur.

2.3 CCM P-sample nonmatches

The 2010 Census Coverage Measurement Program (CCM) evaluated the coverage of the 2010 Census using the PES methodology and provided some data about the coverage of children ages 0 to 4 in the census. The implementation used a sample of census enumerations, known as the E sample, and a sample of the population selected independently from the census, known as the P sample. The samples were drawn by first selecting a sample of block clusters and then taking all the census enumerations coded to the sample block clusters as the E sample and collecting an independent list of the population in the same block clusters for the P sample.

As part of the operation, all the enumerations in the E sample that had sufficient information to identify the person uniquely, namely a name and two characteristics, received a code of correct enumeration, erroneous enumeration, or unresolved. Then the P sample was matched to the census and each person in the P sample is coded as matching a census record, not matching a census records, or having an unresolved status. The census enumerations that did not have sufficient information to identify a person uniquely were not eligible for matching. P-sample codes also indicated whether the P-sample housing unit matched a housing unit on the Census Master Address File (MAF) and the overall match status of the household members by indicating whether some or all household members received a status of nonmatch. Each person in the P-sample was matched to the census at his/her Census Day (April 1) address. Therefore, people who moved into their block between April 1 and their P-sample interview day in July or August, called inmovers, were asked for their Census Day address, which was used in matching. Nonmovers, people who did not move between Census Day and the P-sample interview day, were matched in the sample block. The results from the coding of the E sample and P sample were used in a dual system estimator to form estimates of the population size that can be compared to census counts to evaluate the coverage of the census.

The regression model in Section 3.1 uses the P-sample weighted nonmatching young children as a surrogate for young children missed by the census (Moldoff 2008). We would like to be able to identify the children ages 0 to 4 in the P-sample that were missed by the census but that is not possible. However, a nonmatching person in the P sample indicates an enumeration problem because it indicates one of the three things occurred: the person was missed by the census, counted in the wrong place, or had an enumeration that did not have sufficient information to identify the person uniquely. Therefore, the weighted nonmatching children ages 0 to 4 are as close as we can get to a national-level estimate of missed young children. The P-sample nonmatching children ages 0 to 4 include both inmovers and nonmovers. A report from the U.S. Census Bureau (2017c) contains tabulations of the P-sample nonmatching children ages 0 to 4 by selected demographic, housing, and operational characteristics. In the report, Table 6 shows that a weighted 24% of the nonmatching children ages 0 to 4 were inmovers.

2.4 PDB and Low Response Score

For an indicator of whether an area tends to have low response to censuses and surveys, we use the Low Response Score (LRS) that was developed in recent years at the Census Bureau (Erdman and Bates, 2017). The LRS is based on a regression model that predicts the area's mail return rate observed in the 2010 Census. The model uses 25 independent variables to predict an area's LRS[4]. The way to interpret the LRS is that areas with a high LRS tend to have low response to censuses and surveys. The LRS is available on the Census Bureau's Response Outreach Area Mapper (https://www.census.gov/roam) and Planning Database (https://www.census.gov/research/data/planning_database/) and updated with each issue of the Planning Database.

Our analysis uses the LRS based on 2010 – 2014 ACS data found on the 2016 Planning Database (U.S. Census Bureau 2017). Housing units were assigned the LRS associated with the tract of their address. Then the mean LRS for a segment is calculated using all the occupied addresses in the segment. In effect, the mean LRS is the weighted average of the LRS for the tracts in the segment where the weight for a tract is determined by the proportion of the segment's occupied housing unit population that is in the tract.

2.5 Strategy

Our study focuses on using regression modeling to answer two sets of research questions:

1) Do lifestyle segments provide insight about whether areas with the highest numbers of positive responses to the child-specific coverage probes are also the areas with the highest numbers of CCM P-sample weighted nonmatching children ages 0 to 4? If so, is the relationship strong enough to predict the areas with the highest numbers of CCM P-sample nonmatching young children?

2) Do lifestyle segments provide insight about whether areas with the greatest numbers of positive responses to the child-specific coverage probes also the areas least likely to self-respond? Are the positive responses to the child-coverage question in areas less likely to self-respond, or do some areas with average or high self-response have residents uncertain about whether to include a child in their household as living at their address on Census Day?

To answer the first research question, we use the CCM P-sample weighted nonmatching children ages 0 to 4 as a surrogate for missed young children. The strategy involves

[4] The 25 variables included % age 5-17; % age 18-24; % not High School graduate; % below poverty; % female headed households, no husband; mean number persons per household; % related child under age 6; % moved in last year; % renter occupied; % vacant houses; % Black; % married family households; median household income; median house value; % Hispanic; population density; % non-Hispanic White; % age 65+; % males; % college graduates; % moved in last 5 years; % single family units; % single person households; % age 25-44; % age 45-64.

partitioning the tracts in the U.S. by the Tapestry segments and then merging the number of positive responses to the 2010 Census child-specific coverage probes and the CCM P-sample weighted nonmatching children ages 0 to 4 in the segments. Then we fit a regression model of the CCM P-sample weighted nonmatching children ages 0 to 4 on the number of child-specific coverage probes in the segments. Although the household was the unit of data collection, having the unit of analysis in the regression be the segment is intentional to make inferences about the types of neighborhoods where the 2010 Census missed young children. The segment level is useful in planning many aspects of census communications and data collection. In addition, partitioning the 2010 Census and CCM P-sample data by Tapestry segments enables us to examine whether respondents' behavior with respect to the child-specific coverage probes varies by segment. Inferences at the household-level or person-level from a model at the segment level would run the risk of an ecological fallacy.

For the second research question, we explore whether the segments reveal a relationship between the rate of positive responses to the child-specific coverage probes and hard-to-count status as indicated by the mean LRS. This relationship is of interest because the mean LRS is available nationwide prior to the 2020 Census, and therefore, could be helpful in planning the field operations and the communications campaign for the 2020 Census. This model complements the model explored in answering the first research question.

Section 4 describes related work that uses individual census enumerations in the E-sample as the unit of analysis in a model to predict the probability that a census enumeration is in the correct location. These results add support to the results in Section 3 of a relationship between positive responses to the coverage probes and census coverage error.

3. Results

3.1 Regression of CCM P-sample nonmatches on positive responses to coverage probes

When we partitioned both the CCM nonmatching young children and the positive responses to the child-specific coverage probes by the 67 Tapestry segments, we found the correlation coefficient between them to be 0.86. The high correlation led us to use a simple regression model to further explore the relationship between the CCM P-sample nonmatching young children and the positive responses to the child-specific coverage probes. Figure 4 shows the regression model of the 2010 CCM weighted nonmatching children ages 0 to 4 on the number of positive responses to child-specific probes. The estimated intercept is -5,864 with a standard error of 2,932 and estimated coefficient is 4.03 with a standard error of 0.29, which indicates that both are statistically significant. The regression model has a good fit as indicated by the r-square of 0.75 and an acceptable residual pattern.

We also examined a comparable regression model using rates to check whether the main driver of the correlation was the number of positive child-specific coverage probes in segments merely reflecting the size of the segments, which could be the case if the rate was uniform across segments. Although the correlation between the weighted nonmatch rate and number of positive probes per thousand was 0.58, the r-square of a regression of the weighted nonmatch rate on the number of positive probes per thousand was 0.34. One reason for the low r-square was an apparent outlier segment that had a weighted nonmatch rate three times its unweighted nonmatch rate. When the regression was run a second time without the outlier segment, the correlation increased to 0.67 and the r-square increased to 0.45. The weights evidently contribute more variability to the nonmatch rates than the levels, which results in the r-square for a model using rates to be 0.3 lower than observed for model using levels. The model using levels is better for prediction.

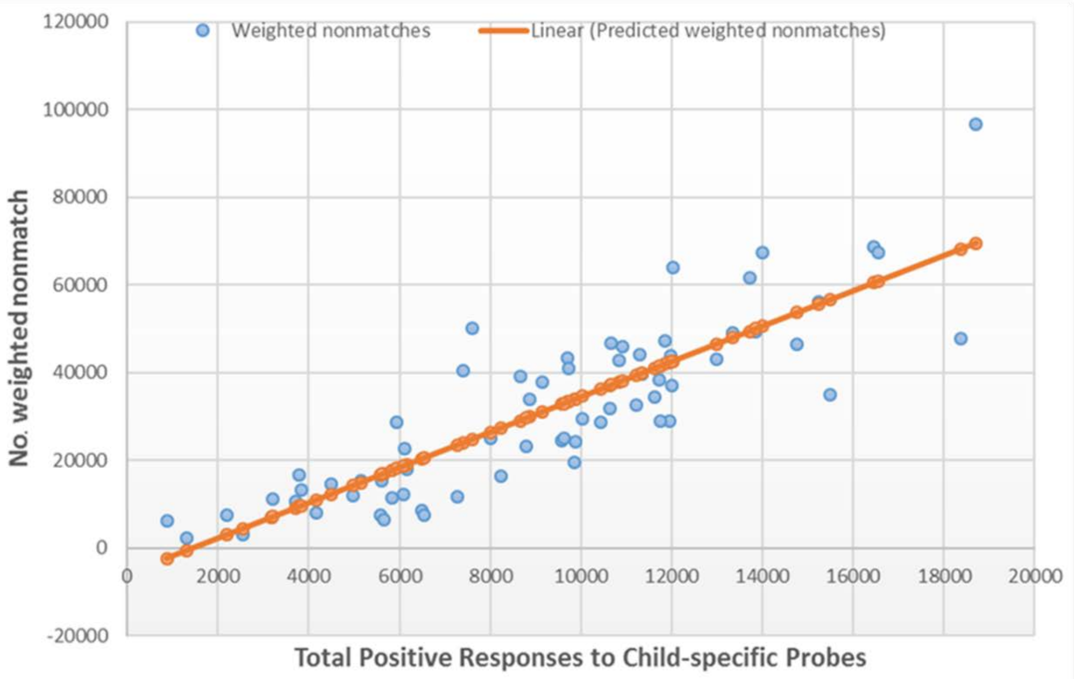


Figure 4. Regression of weighted CCM P-sample Nonmatching Young Children on total Positive Responses to Child-specific Probes in the 2010 Census for Tapestry segments. Note: number of segments = 67; r-squared = 0.75

The regression model in Figure 4. shows that when partitioned by the 67 Tapestry segments, the number of positive responses to child-specific coverage probes capture a large amount (75%) of the variation in the weighted CCM nonmatching young children. In addition, the result indicates that the variable defined by the positive responses to child-specific coverage probes is a good predictor of the weighted CCM nonmatching young children at the segment level.

3.2 Regression of positive responses to coverage probes on mean LRS

When we turned to examine the relationship between the LRS and the positive responses to child-specific coverage probes, we had nationwide data from census questionnaires. Initially, we tried to fit a simple regression model of the number of positive responses to the child-specific per thousand households in a segment on the mean LRS in a segment. The result produced several low outliers and was not a good fit. We noticed that the outlier segments tended to be in the same two LifeModes and the all the observations in the two LifeModes were below the regression line.

One of these two LifeModes was Scholars and Patriots and the other was Uptown Individuals. As shown in Table 1, Scholars & Patriots and Uptown Individuals are the LifeModes with the fewest number of children. Both Scholars & Patriots and Uptown Individuals have a high percentage of single households and adults-only households. Scholars & Patriots tend to be under age 25 and Uptown Individuals are a little older but under age 35. These characteristics led us to speculate that the households in the segments contained in these two LifeModes may have a response pattern for the child-specific coverage probes that is different from the pattern in the other segments. For example, when children are present in households in these segments, respondents may tend to be more certain about where to count them than respondents in similar situations in other segments.

To investigate further, we defined a two-level categorical variable where one level was the combination of two LifeModes Scholars and Patriots and Uptown Individuals and the second level was the 12 other LifeModes combined. Then we used SAS Procedure GLM to fit a multiple regression model for the number of positive probes per thousand households on the mean LRS, the two-level categorical variable and their interaction. Figure 5 shows the ANCOVA plot for this model, which has an r-square equal to 0.83. The F statistics for the model and the Type I and Type II sum of squares for all the variables are statistically significant and a 5-fold cross validation showed that the estimates of the parameters are stable. In the model for the Other LifeModes, the estimate of the intercept is statistically significant implying it is different from zero. In addition, the slope of the line, as indicated by the coefficient of the mean LRS, is statistically significant different from zero. However, in the model for the combination of the LifeMode Scholars & Patriots and the LifeMode Uptown Individuals, the estimated intercept is not statistically significant from zero. A test of the hypothesis that difference in the slopes of the two lines is zero failed, which indicates the two lines are not parallel.

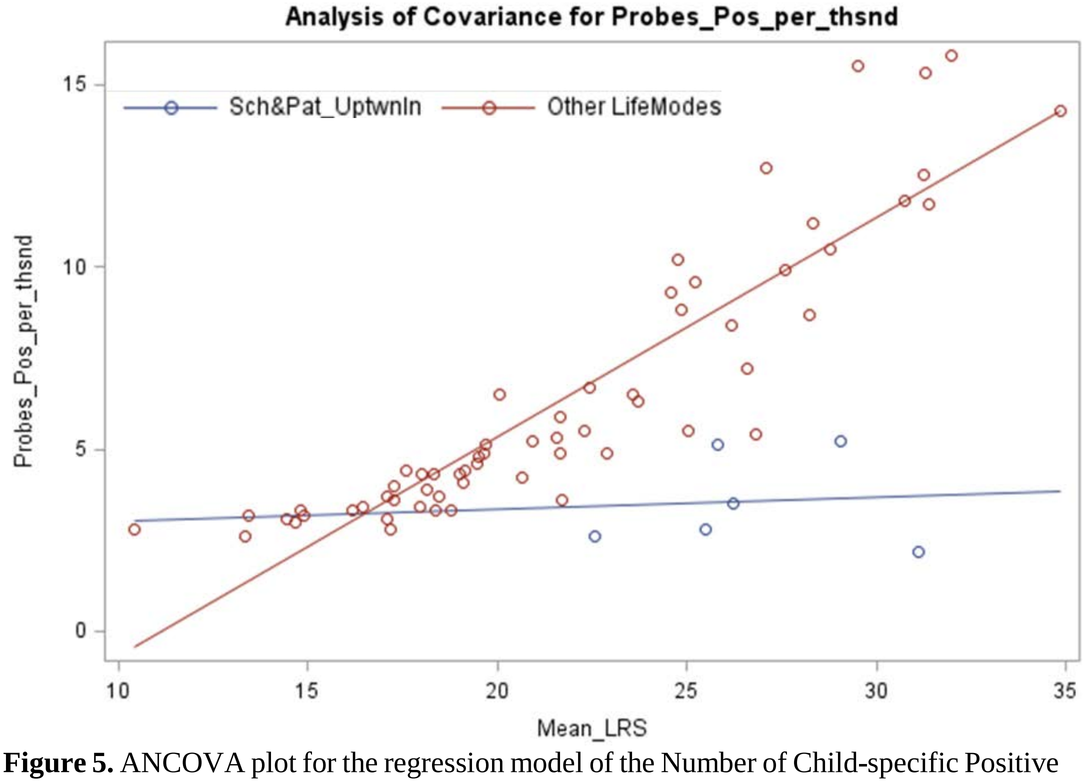


Figure 5. ANCOVA plot for the regression model of the Number of Child-specific Positive Probes per Thousand Households in segments on Mean LRS, LifeModes collapsed to 2 levels, and their interaction. Note: number of segments = 67; R-squared = 0.83.

We also fit a regression model using the logarithmic transformation of the number of positive responses to child-specific probes per thousand households as the dependent variable with the same independent variables as the model in Figure 5. The residuals are better behaved, the estimated coefficients are significant and stable in a 5-fold cross validation, and the r-square equaled 0.84, a little higher than the r-square for the model in Figure 5. However, since the alternative model used a transformation on the dependent variable, predictions on the desired scale would not only require a reverse transformation but also need a bias correction. The model in Figure 5 is better for prediction.

The regression results show that the mean LRS for a segment is a good predictor of the number of positive responses to child-specific coverage probes per thousand households for 12 of the LifeModes. The finding for these 12 LifeModes is interesting because the LRS was designed to predict self-response as opposed to response to coverage probes or other types of questions. In the remaining two LifeModes, the Scholars and Patriots and the Uptown Individuals, the mean LRS for segments appears to be unrelated to the number of positive child-specific probes per thousand households.

4. Related work

The coverage probes were a new feature of the 2010 Census so little research is available on the topic of whether positive responses to the coverage probes provide a source for prediction of coverage problems in the Census. The model in Figure 4 provides some evidence that positive responses for the child-specific coverage probes are predictive of CCM P-sample nonmatching young children when partitioned by the Tapestry segments. However, more research is needed.

There is one other source of analyses that provides some evidence that positive responses to the coverage questions are related to coverage. Giffin (2018) used 2010 E-sample data to fit logistic regression models to estimate the probability of a census enumeration being correct in its block cluster search area, which is defined as the block cluster and the surrounding adjacent blocks. For an enumeration to be correct, it had to be in the right location, which is where the person lives or stays on Census Day, April 1, 2010. The design of the search area built in a small tolerance on the location of a person's residence that made the PES results robust to minor geocoding errors, such as a housing unit being geocoded to the wrong side of the street.

Griffin (2018) used the 2010 E sample, approximately 300,000 person enumerations, to fit the logistic regression model. The independent variables were the ones used in the same type of model to produce the small area estimates of 2010 Census coverage error (Keller and Fox 2012). The variables reflected the following demographic, geographic, and census operational variables:

- Race/Hispanic Origin domains
- Tenure
- Age/Sex groups
- Region of the country
- Metropolitan Statistical Area Size by Type of Enumeration Area
- Presence of Spouse in Household
- Relationship to Householder
- Tract-level Census Participation Rates
- Bilingual and Replacement Questionnaire Mailing Areas

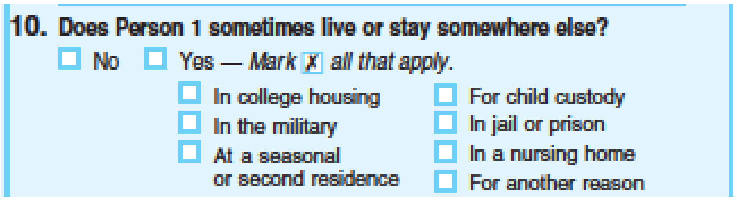
Griffin's model uses two additional variables that were based on questions on the 2010 Census questionnaires. Both of these variables are relevant to our research because they concern census coverage. We discuss them further below.

One of the variables was based on the question containing the coverage probes in Figure 2 that we used in our research. It is sometimes called the undercount question. Our models in the previous section used only the positive responses to the probe that asked whether any children were not included in the count of the number of people living or staying at the address. In Griffin's binary variable called addpeople, one level (addpeople = 0) represents a positive response to any of the categories for possible household members, and the other level (addpeople = 1) reflects the response that no other people were living at the address.

The addpeople variable was significant in the logistic regression model for the probability of being correctly enumerated in the search area of the block cluster. The odds of being a correct enumeration in the block cluster search area if a person living there was not included in the population count (addpeople = 0) divided by the odds for being a correct enumeration in the block cluster if all persons living there were included in the population count (addpeople = 1) was 0.525. Therefore, when the response indicated no other people were living at the address (addpeople = 1), the odds of an enumeration in the housing unit being correct in the block cluster search area are about *twice* the odds of an enumeration in the housing unit being a correct when a person living there was not included in the population count (addpeople = 0).

The odds ratio result is interesting because it applies to all the enumerations in the household. The implication of the estimated odds ratio is that when one or more persons living at the residence were not included in the population count provided by the respondent, the people who were included in the count are less likely to be correctly enumerated. When a respondent is uncertain about whether to include a person staying at the address in the population count, the implication may be that the household membership fluctuates, which could make it difficult determine household membership on a given day.

Griffin's other binary variable was based on the answer to a different question, one sometimes called the overcount question, that was collected for each person and may be viewed in Figure 6. After questions asking a person's name, age and other characteristics, the overcount question asked if the person sometimes lives or stays at another place. One level of the variable, called mayelse, reflected the answer that the person lives elsewhere sometimes (mayelse = 0) and the other level represented the answer that the person does not live elsewhere sometimes (mayelse = 1). This variable was significant in the logistic regression model. The odds of the person's enumeration being correct when the person lives elsewhere sometimes (mayelse = 0) divided by the odds of the person's enumeration being correct when the person does not live elsewhere sometimes (mayelse = 1) is 0.175. Therefore, when the person does not live elsewhere sometimes (mayelse equals 1), the odds of the person's enumeration being correct in the block cluster search area are about 5.7 times the odds of a person's enumeration being a correct when the person lives elsewhere sometimes (mayelse = 0).



10. Does Person 1 sometimes live or stay somewhere else?

□ No □ Yes — *Mark* ☒ *all that apply.*

□ In college housing
□ In the military
□ At a seasonal or second residence
□ For child custody
□ In jail or prison
□ In a nursing home
□ For another reason

Figure 6. Facsimile of Self-Response Questionnaire's Overcount Question for the 2010 Census

The result indicates that positive responses to the question asking whether a person sometimes lives or stays elsewhere contains valuable information about which enumerations may not be in the correct location or may be duplicates of other enumerations.

Griffin's results and our results concerning the relationship between the coverage probes and census coverage error are complementary. Griffin used positive responses to coverage probes for all four types of household members while our analyses used only positive responses to the child-specific coverage probes. The variables in Griffin's model were chosen because they were correlated with census coverage error and therefore, provide a valuable description of households that have relatively high probabilities of coverage errors. While undoubtedly there is a great deal of overlap in the geographic and demographic variables used in Griffin's model and variables used in forming the Tapestry segments, the Tapestry segments also use purchasing patterns and attitudinal information that Griffin's model does not include. The combination of information provided by Tapestry is intended for use in commercial marketing and also may be helpful in designing and monitoring fieldwork in censuses and surveys.

5. Summary

In this study, we focused on analyses regarding the undercount of young children ages 0 to 4 that has been observed in several previous U.S. Censuses. One of our goals was to explore whether lifestyle segments provided insight about the relationship between the positive responses to the child-specific coverage probes and weighted nonmatching young children observed in the 2010 CCM P sample. Our second goal was to investigate whether the areas with the greatest numbers of positive responses to the child-specific coverage probes are also the areas least likely to self-respond. The presented investigations used simple and multiple regression modeling.

For the first goal, we used a simple regression model to explore the relationship between of the positive responses to child-specific probes on the 2010 Census questionnaire and 2010 CCM P-sample weighted nonmatching children ages 0 to 4. Our research implies that the number of positive responses to child-specific probes is predictive of the weighted number of P-sample nonmatching young children ages 0 to 4 when the data are partitioned by the Tapestry segments. The slope of the regression line was positive which indicates that the number of P-sample nonmatching young children increases as the number of positive responses increases. In addition, the relationship appears strong enough to suggest that the areas with the highest numbers of positive responses to the child-specific coverage probes predict areas where CCM P-sample found the highest numbers of nonmatching young children.

With respect to the second goal, our analyses used a multiple regression model to investigate the relationship between number of positive responses to the child-specific coverage probes per thousand households and the mean LRS in segments. The segments in two LifeModes appeared to be a subpopulation that systematically had lower rates for positive responses to the child-specific probes than the rest of the segments with similar values of the mean LRS. Because of this, we included a two-level categorical independent variable, defining one level to comprise both the LifeMode Scholars and Patriots and the LifeMode Uptown Individuals and the second level to cover the other 12 LifeModes. We also included the interaction between the mean LRS and the two-level categorical variable.

The regression model produced two regression lines, one for each level of the categorical variable that had its own slope and own intercept. For the line for the 12 LifeModes, both estimates for the intercept and the slope were statistically significant. However, the intercept for line for the other two LifeModes was not significant and a statistical test comparing its slope to the slope of the line for the 12 LifeModes found that they were different.

The results of the regression modeling indicate that for segments in the 12 LifeModes, the mean LRS in a segment was predictive of the number of positive responses to child specific probes per thousand households for segments. This result may be interpreted as indicating that the segments that tend to have low response rates in censuses and surveys also tend to be the areas where a high percentage of respondents were uncertain about whether to include one or more young children in their household census population count. However, there appears to be a subpopulation composed of the segments in two LifeModes where the rates of positive responses to the child-specific coverage probes are uncorrelated with the mean LRS, and the observed rates tend to be lower than would be expected for the rest of the segments with comparable values of their mean LRS. These LifeModes, Scholars and Patriots and Uptown Individuals, are the two with the fewest number of children in the Census and are among the youngest. Both LifeModes have a high percentage of single-person households and adults-only households. Scholars & Patriots tend to be under age 25 and Uptown Individuals are a little older but under age 35.

The finding that the number of positive responses to the child-specific coverage probe is predictive of census coverage error of children ages 0 to 4 is somewhat validated in a study by Giffin (2018). Griffin fit a logistic regression model to predict the probability of persons enumerated on 2010 Census mail returns being enumerated in the search area of their housing unit, which was defined as its block cluster and surrounding blocks. The model applied to responses for all the coverage probes and therefore to persons of all ages, not just children ages 0 to 4. When the respondent indicated all persons living at the address were included in the population count, the odds of an enumeration in the housing unit being correct in the block cluster search area was about *twice* the odds of being correct when the answer indicated the population count did not include one or more people living at the address.

More research is needed to corroborate our findings, to gain a better understanding of how and why the undercount of young children occurred and to determine how the Tapestry segmentation may contribute to improving the census count for young children. For example, the Coverage Followup (CFU) attempted to contact households that gave a positive response to the child-specific coverage probe. However, the segments with large numbers of positive responses, where one would think the CFU corrected errors, proved to be the segments with a large number of weighted nonmatching young children. This may have happened because the CFU did not resolve the problems, or CFU resolved some of the problems, but was unable to make contact with many of the households. Another possible explanation is that the CFU corrected errors for the specific types of problems, but many other similar problems in these segments were not detected by the probes. The Tapestry segments may aid in identifying the reasons for the variation in the positive responses to the child-specific probes and the causes for the undercount of children ages 0 to 4 if they are lifestyle related.

One line of future research may be to explore models that use P-sample housing units or persons as the unit of analysis. These models would have their own set of challenges since many of the households with P-sample nonmatching children ages 0 to 4 did not have a corresponding census questionnaire at their Census Day address. Therefore, not every nonmatching young child was in a household that had an opportunity to answer the child-specific coverage probes or enumerate the child. The household may not have received a census questionnaire because its address was not on the MAF, as was the case for 15.5 percent (weighted) of the nonmatching children ages 0 to 4 in nonmover households (U.S. Census Bureau 2017c, Table 8). Other times, the household's address was on the MAF but the census status was vacant, as was the case for 11 percent (unweighted) of the nonmover nonmatching young children and 18 percent (unweighted) of the inmover nonmatching young children (U.S. Census Bureau 2017c, Table 9).

The proposed 2020 Census questionnaire includes the undercount question with the coverage probes and overcount question on whether a person sometimes lives or stays elsewhere (U.S. Census Bureau 2018, p. 19). The design of the automated instruments for collecting census information during self-response, Nonresponse Followup and Census Questionnaire Assistance allow for collecting the names of the household members whom the respondent did not include in the population count. When the respondent gives a positive response to a coverage probe, the instrument immediately presents a screen requesting that the respondent enter the names of those excluded. Therefore, the preparations for 2020 Census present an opportunity for designing interventions in the form of messages in the communications campaign and fieldwork to improve the undercount of young children. The results of the regression models may be helpful in the design and planning of operational aspects of the interventions since the segments and the LRS are available on a nationwide basis prior to the Census. During the Census data collection, positive responses to the child-specific coverage probes may aid in implementing targeted re-contacts or decisions on whether to add children found in administrative records at the address to the household. During the 2020 PES processing, positive responses to coverage probes may provide some evidence that is helpful in confirming that two enumerations that appear to be for the same person really are for the same person.

The distribution of positive responses to the coverage probes and the overcount question on the 2010 Census questionnaire may be helpful in planning the 2020 Census, the 2020 PES, and future census-taking and coverage evaluations. Models fit by Griffin (2018) were intended for use in small area estimation of census coverage error but current plans for the 2020 PES do not include producing these estimates. However, the result that the coverage probes and the overcount question are influential variables in the model to predict the probability that a census enumeration is in the correct location may be helpful in other ways. Other operations in the 2020 Census and the 2020 PES may be able to leverage the relationship between positive responses to the coverage probes and P-sample nonmatches and the variation in the number of positive responses to the child-specific coverage probes found in Tapestry segments. Further research would be helpful in determining how to use the coverage probes and overcount question to improve census-taking.

References

Coale, A. J. 1955. The population of the United States in 1950 classified by age, sex, and color-a revision of census figures. *Journal of the American Statistical Association*, 16-54.

Erdman, C. and N. Bates. 2017. "The Low Response Score: A Metric to Locate, Predict, and Manage Hard-to-Survey Populations," *Public Opinion Quarterly*, 81 (1), 144-156. DOI: https://doi.org/10.1093/poq/nfw040.

Esri. 2017. “Tapestry Fliers,” Redlands, CA: ESRI. Available at http://doc.arcgis.com/en/esri-demographics/data/tapestry-segmentation.htm#ESRI_SECTION1_87F5D845F8E04723AE1F4F502FF3B636

Experian. 2018. “ConsumerView: Marketing Data that Connects Brands with Fans.” Costa Mesa, CA: Experian. Available at http://www.experian.com/marketing-services/targeting/data-driven-marketing/consumer-view-data.html

Griffin, D. and S. Konicki, 2017a. Investigating the 2010 Undercount of Young Children - Geographic Analysis of Coverage Followup Results. Decennial Statistical Studies Division, U.S. Census Bureau. Washington, DC: U.S. Census Bureau. Available at https://www2.census.gov/programs-surveys/decennial/2020/program-management/final-analysis-reports/2020-report-2010-undercount-children-coverage-followup-results.pdf

Griffin, D. and S. Konicki. 2017b. Investigating the 2010 Undercount of Young Children – Analysis of Coverage Followup Results Using the Esri Tapestry Segmentation and the Planning Database. Decennial Statistical Studies Division, U.S. Census Bureau. Washington, DC: U.S. Census Bureau. Available at https://www2.census.gov/programs-surveys/decennial/2020/program-management/final-analysis-reports/2020-report-2010-undercount-children-coverage-followup-analysis.pdf

Griffin, D. and S. Konicki. 2017c. Investigating the 2010 Undercount of Young Children – A Comparison of Demographic, Social, and Economic Characteristics of Children by Age. Decennial Statistical Studies Division, U.S. Census Bureau. Washington, DC: U.S. Census Bureau. Available at https://www2.census.gov/programs-surveys/decennial/2020/program-management/final-analysis-reports/2020-report-2010-undercount-children-characterisitcs-by-age.pdf

Griffin, R. 2018. 2020 Coverage Measurement Research Report: Results of Logistic Regression Modeling for Component Error Synthetic Estimates for the 2010 Census Coverage Measurement. DSSD 2020 Post-Enumeration Survey Memorandum Series. Unpublished memorandum. Decennial Statistical Studies Division, U.S Census Bureau. Washington, DC: U.S. Census Bureau.

Jarmin, R. 2018. Improving Our Count of Young Children. Director’s Blog post July 2, 2018. U.S. Census Bureau. Washington, DC: U.S. Census Bureau. Available at https://www.census.gov/newsroom/blogs/director/2018/07/improving_our_count.html

Keller, A. and T. Fox. 2012. “2010 Census Coverage Measurement Estimation Report: Components of Census Coverage for the Household Population in the United States”, DSSD 2010 Census Coverage Measurement Memorandum Series #2010-G-04. Decennial Statistical Studies Division. Washington, DC: U.S. Census Bureau. Available at https://www.census.gov/coverage_measurement/pdfs/g04.pdf

Konicki, S. and D. Griffin. 2018. Summary of Recent Research on the Undercount of Young Children in the Decennial Census. Unpublished manuscript, draft. Decennial Statistical Studies Division. Washington, DC: U.S. Census Bureau.

Moldoff, M. 2008. “The Design of the Coverage Measurement Program for the 2010 Census” DSSD 2010 Census Coverage Measurement Memorandum Series #2010-B-7. Decennial Statistical Studies Division. Washington, DC: U.S. Census Bureau. Available at 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. Decennial Statistical Studies Division. Washington, DC: U.S. Census Bureau. Available at https://www.census.gov/coverage_measurement/pdfs/g01.pdf

Mulry, M., N. Bates, and M. Virgile. 2018. Viewing Participation in Censuses and Surveys through the Lens of Lifestyle Segments. Unpublished manuscript. Center for Statistical Research and Methodology, U.S. Census Bureau. Washington, DC: U.S. Census Bureau.

O'Hare, W. P. 2015. *The undercount of young children in the US Decennial Census*. New York, NY: Springer International Publishing.

Siegel, J. and M. Zelnik. 1966. An evaluation of coverage in the 1960 census of population by techniques of demographic analysis and by composite methods. *Proceedings of the Social Statistics Section*. Alexandria, VA: American Statistical Association, 71-85.

U.S. Census Bureau. 2018. Questions Planned for the 2020 Census and American Community Survey. Washington, DC: U.S. Census Bureau. Available at https://www2.census.gov/library/publications/decennial/2020/operations/planned-questions-2020-acs.pdf

U.S. Census Bureau. 2017a. 2016 Planning Database. Washington, DC: U.S. Census Bureau. Available at https://www.census.gov/research/data/planning_database/2016/

U.S. Census Bureau. 2017b. Investigating the 2010 Undercount of Young Children – Analysis of Census Coverage Measurement Results. Washington, DC: U.S. Census Bureau. Available at https://www2.census.gov/programs-surveys/decennial/2020/program-management/final-analysis-reports/2020-2017_04-undercount-children-analysis-coverage.pdf

U.S. Census Bureau. 2012. Documentation for the Revised 2010 Demographic Analysis Middle Series Estimates. Population Division. Washington, DC: U.S. Census Bureau. Available at http://www.census.gov/popest/research/DA_Methodology.pdf

U.S. Census Bureau. 2010. The Development and Sensitivity Analysis of the 2010 Demographic Analysis Estimates. The Demographic Analysis Research Team, Population Division. Washington, DC: U.S. Census Bureau. Available at http://www.census.gov/newsroom/releases/pdf/20101206_da_revpaper.pdf

West, K., and J. G. Robinson. 1999. What Do We Know about the Undercount of Children? Population Division. Washington, DC: U.S. Census Bureau. Available at http://www.census.gov/library/working-papers/1999/demo/POP-twps0039.html