The Sample Design of the U.S. Census Bureau's 2015 National Content Test[1]

Sarah Konya, Kelly Mathews, Michael Bentley
U.S. Census Bureau, 4600 Silver Hill Road, Suitland, MD 20746

Abstract
The 2015 National Content Test (NCT) is the primary mid-decade opportunity to test content on a nationally representative sample as part of the research and development cycle leading up to the re-engineered 2020 Census. In addition to testing content modifications, the 2015 NCT also tested different contact strategies designed to optimize self-response and different approaches to offering in-language materials. The 2015 NCT employed a complex sample design to select 1.18 million stateside housing unit addresses and 20,000 Puerto Rico housing unit addresses. The work presented describes the coverage, race, and response propensity strata that were assigned, as well as how the 1.18 million stateside housing unit address sample was selected using these strata.

Key Words: Census test, sample design

1. Introduction

The 2015 National Content Test (NCT) is the primary mid-decade opportunity to test content on a nationally representative sample as part of the research and development cycle leading up to the re-engineered 2020 Census. Content testing for the 2015 NCT included race and ethnicity, relationship, and within-household coverage. In addition to testing content modifications, the 2015 NCT also tested different contact strategies designed to optimize self-response and different approaches to offering in-language materials.

The 2015 NCT included a test in the San Juan Municipio of Puerto Rico, which is the most urban area of Puerto Rico. The 2015 NCT was the first census test in which the Census Bureau mailed materials to housing unit addresses in Puerto Rico. Before the 2015 NCT, enumerators would visit housing unit addresses in Puerto Rico and enumerate the household at the time of their visit or would leave a form for the household to fill out and send back.

Census Day for the 2015 NCT was September 1, 2015, which is the day that the Census Bureau asks the respondent to reference when listing the people who were living at that address. A sample of 1.18 million stateside housing unit addresses and 20,000 Puerto Rico housing unit addresses were selected to participate in the 2015 NCT. Along with the self-response portion of the test, a subsample of housing unit addresses was selected to participate in the race and ethnicity and coverage reinterviews, which were designed to measure the accuracy of the collection of race and ethnicity and coverage data, respectively.

[1] Disclaimer: This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

2. Methodology for the Stateside Sample

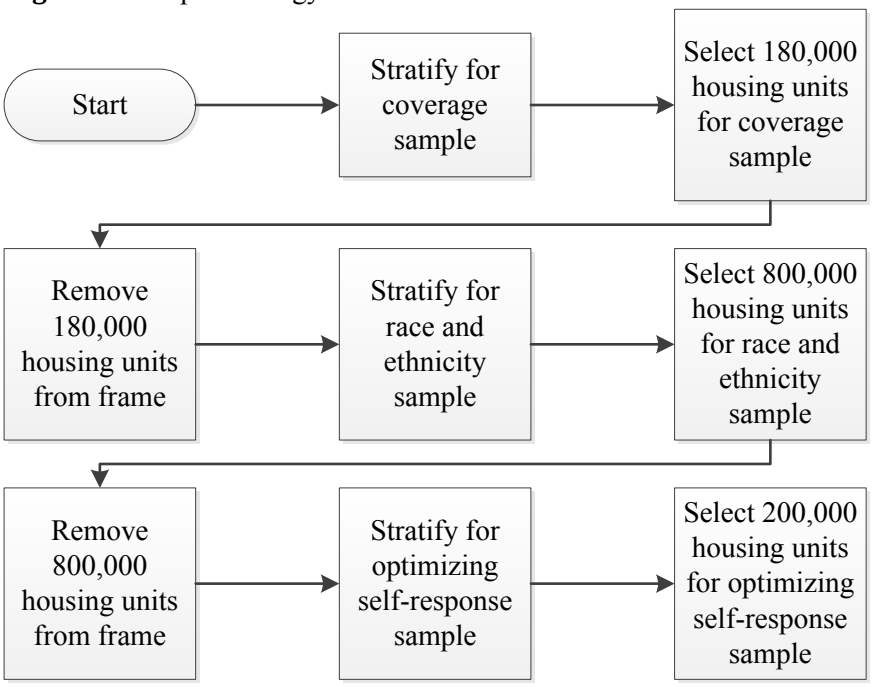
2.1 Survey Frame

The survey frame was the Census Bureau address list as of January 2015, which contained 138,648,951 housing unit addresses. Group quarters and addresses deemed to be unmailable were excluded from the frame. In order to reduce respondent burden, housing unit addresses that had been selected to participate in the 2015 Census Test or the 2015 American Community Survey (ACS) were also excluded from the frame. The frame contained 132,641,277 housing unit addresses after all exclusions.

2.2 Stratification and Sample Selection

The universe was stratified and sampled for three separate research topics: coverage, race and ethnicity, and optimizing self-response (OSR). The purpose of this approach was to ensure a sufficient sample of households in each of the three research areas. This approach was carried out as Figure 1 below describes.

Figure 1: Sample Strategy for the 2015 NCT



Although three separate samples were selected, all respondent data will be used to collectively conduct the analyses. Sampling weights that reflect the complex sample design were assigned to the in-sample housing unit addresses.

2.2.1 Coverage

The goal of the coverage sample was to reach tracts that were expected to have significant numbers of households susceptible to coverage overcounts. Coverage overcounts are respondents that have indicated that they sometimes live or stay somewhere other than the address to which the mail material (i.e. letter, postcard, questionnaire) was sent. The coverage strata shown in Table 1 were defined by subject matter experts at the Census Bureau before the 2015 NCT sample was selected. Only 937 of the 72,238 tracts in the frame were eligible for the coverage sample. Using five-year

ACS data (2009 – 2013), those 937 tracts were assigned to one of the strata in the order in which the strata are listed in Table 1. As an example, if a tract was eligible for both the Jail and Nursing Home strata, the tract was assigned to the Jail stratum.

Table 1: Coverage Strata

Coverage Stratum	**Definition**
Child Custody	25 percent or more of the occupied housing units had a separated or divorced adult and the presence of a child less than 18 years old
Jail	70 percent or more of the occupied housing units were located in an urban area and met the poverty definition
Military	80 percent or more of the occupied housing units had at least one person in active duty military since September 2001
College	50 percent or more of the occupied housing units had a young person between the ages of 6-26 and a college-educated adult (Bachelor's Degree or higher) between the ages of 35-65
Seasonal	60 percent or more of the occupied housing units had a person between the ages of 45-75 and a household income of at least $100,000
Nursing Home	85 percent or more of the occupied housing units had the presence of someone age 60 or older

After stratification, the housing unit addresses were sorted by state, county, tract, and a unique housing unit identification variable. A stratified systematic sample of 180,000 housing unit addresses was selected. The design effect was taken into consideration when choosing the sample sizes for each stratum displayed in Table 2.

Table 2: Coverage Sample Sizes by Stratum

Stratum	**Number of Housing Unit Addresses in Stratum**	**Sample Size**
Child Custody	63,570	30,000
Jail	110,185	30,000
Military	135,647	30,000
College	227,917	30,000
Seasonal	273,925	30,000
Nursing Home	523,019	30,000

2.2.2 Race and Ethnicity

The 180,000 housing unit addresses selected for the coverage sample were removed from the sampling frame, and the remaining housing unit addresses were eligible for the race and ethnicity sample. The goal of the race and ethnicity sample was to obtain a diverse sample with representation from a variety of race and ethnic groups. Using five-year ACS data (2009 – 2013), tracts were assigned to one of strata in Table 3 in the order in which the strata are listed. Race and ethnicity estimates in each stratum and the stratum size were taken into consideration when setting the thresholds in Table 3.

Table 3: Race and Ethnicity Strata

Race and Ethnicity Stratum	Definition
Middle Eastern or North African (MENA)	10% or more of the population identified as MENA using either race or ancestry data
American Indian or Alaska Native (AIAN)	10% or more the population identified as AIAN
Asian/Native Hawaiian Other Pacific Islander (NHPI)	15% or more of the population identified as Asian or NHPI
Black	25% or more of the population identified as Black
Hispanic	45% or more of the population identified as Hispanic
All Other	All remaining tracts that were not assigned to one of the previous strata

Table 4 displays the ACS estimates for the MENA, AIAN, Asian, NHPI, Black, and Hispanic populations. The yellow cells highlight the race or ethnicity estimate for the respective stratum. The values in each of the yellow cells are relatively large when compared to their respective race or ethnicity estimates for the nation as a whole. Therefore, the strata assignment was successful at targeting those groups. In addition to the race and ethnicity estimates, the 2010 Census mail response rate is shown for each stratum. The AIAN stratum had the lowest 2010 Census mail response rate at 57.8 percent, and the Asian/NHPI stratum had the highest 2010 Census mail response rate at 69.0 percent.

Table 4: Estimates of Race and Ethnicity Strata in the 2015 NCT Universe

Stratum	Number of Housing Units	MENA %	AIAN %	Asian %	NHPI %	Black %	Hisp. %	2010 Mail Resp. Rate*
MENA	1,302,686	19.1	1.0	14.0	0.4	9.4	15.5	67.4
AIAN	1,996,773	0.3	26.3	2.5	0.5	6.9	12.7	57.8
Asian/ NHPI	11,626,285	2.2	1.3	29.2	2.0	9.0	20.0	69.0
Black	20,783,674	0.5	1.1	2.5	0.2	53.5	16.1	59.6
Hisp.	9,474,271	0.6	1.4	3.1	0.3	7.1	73.6	62.1
All Other	87,277,588	0.9	1.2	3.2	0.2	5.7	8.9	68.2

Source: 2009-2013 ACS Data
*Rate calculated from the Planning Database

After stratification, the housing unit addresses were sorted by state, county, tract, and a unique housing unit identification variable. A stratified systematic sample of 800,000 housing unit addresses was selected. The design effect was taken into consideration when choosing the sample sizes for each stratum displayed in Table 5.

Table 5: Race and Ethnicity Sample Sizes by Stratum

Stratum	Number of Housing Unit Addresses in Stratum	Sample Size
MENA	1,302,686	100,000
AIAN	1,996,773	100,000
Asian/NHPI	11,626,285	100,000
Black	20,783,674	160,000
Hispanic	9,474,271	160,000
All Other	87,277,588	180,000

2.2.3 Optimizing Self-Response

The 180,000 housing unit addresses selected for the coverage sample and the 800,000 housing unit addresses selected for the race and ethnicity sample were removed from the sampling frame, and the remaining housing unit addresses were eligible for the OSR sample. The goal of the OSR sample was to obtain households with various levels of response propensity in order to test the experimental contact strategies. The OSR strata were created using two data elements: the Low Response Score (LRS) from the Planning Database (PDB) and Internet access data from the Federal Communications Commission (FCC).

The LRS came from the PDB, which is a tract-level database of housing, demographic, socioeconomic, and operational census variables extracted from the 2010 Decennial Census and the 2008 – 2012 ACS databases. The LRS is the Census Bureau's updated, model-based hard-to-count score developed after the 2010 Census. The LRS is a continuous score that predicts whether a tract will produce a low mail return rate. The score is inversely related to the mail return rate of the 2010 Census for that tract (Erdman and Bates, 2014). Therefore, tracts with a lower LRS are more likely to have a mail return than those with a higher LRS. The measures of spread of the LRS were found and used to create three response categories: Low, Medium, and High. Table 6 provides the LRS categories used for sampling.

Table 6: Low Response Score and Categories

Low Response Score Range from PDB	Response Category
$x \geq 25$ or x = missing	Low
$20 \leq x < 25$	Medium
$0 \leq x < 20$	High

The FCC collects data biannually concerning subscribership to Internet access services in the fifty states, the District of Columbia, and the U.S. territories of American Samoa, Guam, Northern Mariana Islands, Puerto Rico, and U.S. Virgin Islands. The data are collected from providers of advanced telecommunications capability, such as telephone companies, cable system operators, and wireless service providers using FCC Form 477 (Wireless Competition Bureau, 2015; Federal Communications Commission, 2015).

The FCC data are a census tract level file that maps Internet access services faster than 200 kbps in at least one direction. The file, current as of December 2013, put each tract into one of six groups indicating a range of the number of residential fixed high-speed Internet connections per 1,000 households. These groups were condensed into the three categories of Internet access shown in Table 7.

Table 7: Tract Groups for FCC Data

Connections per 1,000 Households	Response Category
$0 < x \leq 600$	Low
$600 < x \leq 800$	Medium
$800 < x$	High

In a previous census test, the 2015 Census Test in the Savannah, Georgia Designated Marketing Area, the three LRS categories and three FCC categories combined to create nine strata. Analysis of the response data showed a pattern of low, medium, and high response rates (Mathews and Rothhaas, 2015). Based on that pattern of response, it was decided to use the three strata shown in Table 8 for the 2015 NCT.

Table 8: 2015 OSR Strata

FCC / **LRS**	$0 < x \leq 600$	$600 < x \leq 800$	$800 < x$
$x \geq 25$ or x = missing	**LOW**	**LOW**	**LOW**
$20 \leq x < 25$	**LOW**	MEDIUM	MEDIUM
$0 \leq x < 20$	**LOW**	HIGH	HIGH

Table 9 displays the ACS estimates for the MENA, AIAN, Asian/NHPI, Black, and Hispanic populations and the 2010 Census mail response rates for each stratum. It is clear that the populations differ by stratum. Specifically, the Low and Medium OSR strata have relatively high percentages of Black and Hispanic populations. The 2010 Census mail response rates support the Low, Medium, and High strata assignments.

Table 9. Estimates of OSR Strata in the 2015 NCT Universe

Stratum	Number of Housing Units	MENA %	AIAN %	Asian / NHPI %	Black %	Hisp. %	2010 Mail Resp. Rate*
Low	47,982,302	0.7	1.9	4.6	22.6	28.5	59.5
Medium	27,651,273	1.5	1.7	9.5	14.6	19.7	66.2
High	56,027,702	1.2	1.1	5.4	5.3	6.7	72.2

Source: 2009-2013 ACS Data

*Rate calculated from the Planning Database

After stratification, the housing unit addresses were sorted by state, county, tract, and a unique housing unit identification variable. A stratified systematic sample of 200,000 housing unit addresses was selected. The design effect was taken into consideration when choosing the sample sizes for each stratum displayed in Table 10.

Table 10: OSR Sample Sizes by Stratum

Stratum	Number of Housing Unit Addresses in Stratum	Sample Size
Low	47,982,302	90,000
Medium	27,651,273	50,000
High	56,027,702	60,000

3. Methodology for the Puerto Rico Sample

The sampling frame for the Puerto Rico sample was the census address list for the San Juan Municipio. The frame contained 169,432 housing unit addresses, and 20,000 housing unit addresses were selected using a systematic random sample.

4. Results

The reports with all of the results from the 2015 NCT are forthcoming, but some preliminary OSR analysis produced the response rates and standard errors shown in Table 11. The response rates are shown by mode (Internet, Telephone, and Mail) and overall, and are compared to the 2010 Census mail response rate.

Table 11: Preliminary OSR Results from the 2015 NCT

Stratum	Internet Response Rate	Telephone Response Rate	Mail Response Rate	Overall Response Rate	2010 Mail Resp. Rate
Low	22.5% (0.09)	5.6% (0.04)	11.7% (0.06)	39.9% (0.09)	59.5%
Medium	37.4% (0.14)	5.4% (0.06)	8.7% (0.07)	51.5% (0.13)	66.2%
High	45.8% (0.12)	6.9% (0.06)	9.5% (0.07)	62.2% (0.12)	72.2%

The differences in the 2015 NCT overall response rates by stratum indicate that the OSR strata were successful in identifying low, medium, and high self-response propensity areas. When comparing the 2015 NCT overall response rates to the 2010 Census mail response rates, it is not surprising that they are so different. Census tests typically do not elicit the same level of response as a decennial census due to lack of awareness. It is interesting, however, how different the overall response rates are from the 2010 Census mail response rates by stratum. The Low stratum's overall response rate is 19.6 percentage points lower than the 2010 Census mail response rate, the Medium stratum's overall response rate is 14.7 percentage points lower than the 2010 Census mail response rate, and the High stratum's overall response is 10 percentage points lower than the 2010 Census mail response rate. Additional research is being conducted to improve the identification of areas of the country that have lower propensities to respond.

References

Erdman, C. and Bates, N., (2014), "The U.S. Census Bureau Mail Return Rate Challenge: Crowdsourcing to Develop a Hard-to-Count Score". U.S. Census Bureau: Center for Statistical Research and Methodology Report Series (Statistics #2014-08). Retrieved March 12, 2015 from https://www.census.gov/srd/papers/pdf/rrs2014-08.pdf

Federal Communications Commission. 2015. Broadband Deployment Data from FCC Form 477. Retrieved March 2015 from https://www.fcc.gov/general/broadband-deployment-data-fcc-form-477

Mathews, K. and C. Rothhaas. 2015. Sample Design Specifications for the 2015 Savannah Site Test. DSSD 2020 Decennial Census R&T Memorandum Series #E-05, U.S. Census Bureau.

Wireline Competition Bureau. 2013. Internet Access Services: Status as of December 31, 2012. Retrieved March 2015 from https://apps.fcc.gov/edocs_public/attachmatch/DOC-324884A1.pdf