The American Community Survey Sample Design: An Experimental Springboard

Megha Joshipura, Steven Hefter

U.S. Census Bureau
4600 Silver Hill Road
Washington, D.C., 20233
Megha.P.Joshipura@census.gov
Steven.P.Hefter@census.gov

Introduction

In an ongoing effort to improve the quality of the American Community Survey (ACS) estimates and contain data collection costs, the Census Bureau routinely implements a series of experimental tests including testing of proposed new content, revisions to the current set of ACS questions, and proposed enhancements to data collection methods. This yearly testing program is called the ACS Methods Panel. In 2007, the ACS Methods Panel originally included one experiment designed to assess the effectiveness of obtaining accurate responses to two different versions of a proposed field of degree question on the ACS questionnaire, and to evaluate the difference between asking the basic demographic items (age/date of birth, race, ethnicity, sex, and relationship) in a grid versus a sequential questionnaire design. This experiment, hereinafter called the Original Content Test, was cancelled in December 2006 since the budget for fiscal year 2007 had not yet been determined. Funding was, however, available to cover one experimental component of the Original Content Test, the grid and sequential questionnaire design, with several constraints on the mode of data collection. This test was called the Grid/Sequential Questionnaire Test. Additionally, once funding was received for fiscal year 2007, we were able to reinstate the experimental testing of a new field of degree question as a separate experiment called the 2007 ACS Content Test. So ultimately, the 2007 ACS Methods panel included two tests – the Grid/Sequential Questionnaire Test, and the 2007 ACS Content Test. While there were only two ACS Methods Panel tests conducted in 2007, we designed samples for the two experiments that were fielded in addition to the Original Content Test that was cancelled. This paper describes the sample design for all three experiments.

The requirements of each experiment necessitated similar, yet unique sample designs. The Original Content Test and the 2007 ACS Content test were designed to collect data via mailout, Computer Assisted Telephone Interview (CATI), and Computer Assisted Personal Interview (CAPI) in sequential months. These samples employed several sampling rates, which are proportional to the 2007 ACS sample probabilities of selection. The Grid/Sequential Questionnaire Test was a mail only test. This sample was selected using constant overall sampling rates within each treatment. Each sample consists of 30,000 addresses selected in pairs. Of particular interest is the calculation of the selection probabilities to account for previous samples, including the ACS sample, drawn from the same frame at different points in time.

This paper discusses the unique opportunities provided by the ACS frame, sample design, and selection for experimental sample designs and sampling frames to be derived from it. The added benefit of using the ACS sample, given the readily available response data – which can be used as experimental controls – is also highlighted. For each experimental design we focus on the statistical requirements, including reliability calculations, the selection probability derivations accounting for the differential, two-phase sample design of the ACS, and provide selected results of the samples by treatment.

ACS Sample Design

An independent sample for the ACS is selected for each of the 3,141 counties and county equivalents in the United States, including the District of Columbia, and each of the 78 municipios in Puerto Rico. The ACS sampling frame is derived from the Master Address File (MAF). Each year the ACS samples approximately 3 million housing unit (HU) addresses in the

This report is released to inform interested parties of research and to encourage discussion. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

United States and approximately 36,000 HU addresses in Puerto Rico. The sample is selected in two phases every year, referred to as the main and supplemental phase respectively.

The main phase sample is selected in August/September of the year prior to the sample year from an extract of the MAF delivered at that time. Approximately 99 percent of the total annual ACS sample is selected during main phase sampling and is allocated to each of the twelve monthly panels for the sample year.

In January of the sample year, a sample of addresses that have been added to the MAF since the Main MAF extracts were created is selected. This is known as the Supplemental phase and accounts for approximately one percent of the total ACS sample. These addresses are allocated to the last nine monthly panels of the year.

The sampling for each phase is carried out in two stages. The first stage sample consists of splitting the entire frame into five pieces, each one including approximately 20 percent of the addresses on the frame. These 20 percent samples are maintained over time and units new to the twice-yearly frames are systematically allocated to these five partitions or samples. These samples are rotated each year and the 20 percent sample designated to the sample year becomes the universe for the second-stage of selection. This means each address is only eligible to be selected in sample for the ACS once every five years, minimizing respondent burden. The first-stage sample selection divides the sample into two strata: one for existing addresses and one for new addresses. The first-stage sampling procedure allocates twenty percent of all new addresses to each of the 4 backsamples and the current year's first stage sample. The second-stage sampling procedure subsamples the units selected in the first-stage. Based on the size (number of estimated occupied housing units) of the area the block is in, it is assigned to one of the five second-stage strata employed during this stage. A reduction factor is used in two second-stage strata. The initial sample within these two strata is reduced by 8 percent where they overlap areas expected to have the highest combined mail/CATI cooperation rates. This, in practice, provides seven unique ACS target sampling rates. The target second-stage sampling rates (*R*) used in each second-stage stratum are as follows:

- Base Rate
- 0.92 × Base Rate
- 3 × Base Rate
- 1.5 × Base Rate
- 0.75 × Base Rate
- 0.92 × 0.75 × Base Rate
- 10%

Six of the seven second-stage sampling rates are a function of a base rate, which is recalculated each year using a target sample size of 3 million addresses (U.S.). The second-stage rates are reduced in tracts with high expected mail/CAPI cooperation rates. This was designed to offset the additional cost incurred from the implementation of differential CAPI sampling rates. The CAPI sampling is selected from two categories of cases. Mailable addresses with neither a response to the mailout nor a telephone interview are sampled at a rate of one-in-two, two-in-five, or one-in-three. Unmailable addresses are sampled at a rate of two-in-three. Refer to Asiala (2005) for complete details on the differential CAPI sample design research.

The ACS collects data in three modes: mail, CATI, and CAPI. The mailable addresses selected in the second-stage sample are sent a questionnaire in the mail. Any non-responding addresses with a telephone number are sent to CATI. A subsample of both the non-responding addresses after CATI and the unmailable addresses is sent to CAPI. For detailed information about the ACS program, please see U.S. Census Bureau (2006).

The Original Content Test

The first experiment, the Original Content Test, was designed to assess the effectiveness of obtaining accurate responses to two different versions of a proposed field of degree question on the ACS questionnaire, and to evaluate the difference between asking the basic demographic items in a grid versus sequential method.

This sample design was largely based on the ACS design in order to simulate the conditions under which the new field of degree question may ultimately be asked. Under this design there were both CATI and CAPI follow-up of non-responding addresses. This test did not include Alaska, Hawaii, and Puerto Rico. This test also included a telephone re-interview of the topics of interest such as the field of degree question for all respondents. As the test was to be mailed out in March 2007, the corresponding ACS panel was to serve as the control for this test.

The target universe for the Original Content Test consisted of all valid, residential housing unit addresses in all county and county equivalents in the United States, except Alaska, Hawaii, and Puerto Rico. For more information on how housing unit addresses are deemed valid, see Zimolzak and Rose (2007) and Bates (2006). The sampling frame was the 2007 ACS main phase first-stage sample that was not selected in second-stage sampling or in any other operation (training, other tests, etc.).

In order to determine the target sample size for this test, the minimum detectable differences (MDD) were calculated for the gross difference rate (GDR) for the Field of Degree question for sample sizes of 50,000, 40,000, and 30,000. The following formula was used:

$$MDD = \sqrt{(2PQ)} \times \sqrt{DEFF} \times \left(z_{\alpha/2} - z_{(1-\beta)} \right),$$

where P = E(GDR), Q=(1 – P), $DEFF$ = the complex sample design effect, $z_{\alpha/2}$ = the cutoff point on the standard normal distribution for a significant difference between the GDR for the two treatments, $z_{(1-\beta)}$ = the cutoff point on the standard normal distribution for a power level of at least 80% for this test.

The difference in the MDDs for varying sample sizes showed no appreciable gain with a sample larger than 30,000. The Original Content Test sample design employed the seven ACS second-stage sampling strata with each block assigned to one of the strata.

The experimental design of the test is shown in Table 1 below:

Table 1: Experimental Design of the Original Content Test

Field of Degree Question Version 1		Field of Degree Question Version 2	
Grid Design	Sequential Design	Grid Design	Sequential Design

The requirements of the Original Content Test were as follows:

- Each of the 15,000 addresses in each treatment sample is selected with probability proportional to the ACS sample probability of selection (POS).
- The sample selection should be a paired design. That is, for each selected address, the next address on the list is also selected.
- Each treatment panel has the same number of addresses allocated to it.
- Each pair is not assigned the same questionnaire design or the same field of degree question.

The sample selection procedure used a paired design in which a sample of 15,000 was selected and the nearest available neighbor is also selected into the Original Content Test sample. This helps to ensure that the two samples of 15,000 are similar. Where possible, all pairs of sample records came from the same state/county/ACS second-stage sampling stratum combination. Unbiased sampling weights were also calculated. Each sample of 15,000, when weighted, is representative of the entire universe.

Each selected address was designated to one of four treatment panels based on the cross of the two treatments: field of degree question version and grid or sequential version of the short-form data questionnaire. Each panel had an equal number of sample cases allocated to it. For further detail on sample design, see Joshipura and Hefter (2006).

The Original Content Test sample design was driven by the requirement that each of the 15,000 addresses in each treatment sample be selected with probability proportional to the ACS sample probability of selection (POS). The total number selected for the Original Content Test sample was 30,000. Each design panel was allocated 7,500 addresses. The total number of records allocated to each individual treatment was 15,000. See Joshipura and Hefter (2007a) for additional results.

Derivation of the Selection Probabilities for the Original Content Test
Define the Content Test POS, *P(CT)* within each county /sub-stratum to be:

$$P_{ijk}\left(CT\right)=C\times F_{ijk} \qquad [1]$$

where
C represents a constant scalar used to achieve the desired sample size of 15,000
F_{ijk} represents the factor of P_{ijk} that is proportional to the ACS second-stage POS for the i^{th} county, j^{th} first-stage ACS sampling stratum, and k^{th} ACS second-stage sampling stratum.
i = 1,...,x (x = # of counties available on the Content Test sampling frame).
j = 1,...,y (y = # of first-stage strata in county i)
k = 1,...,z (z = # of second-stage strata in first-stage stratum j

The requirement that the selection probabilities be proportional to the ACS leads to the following necessary condition:

$$F_{ijk}\times\frac{N_{ijk}^{(0)}-n_{ijk}^{(0)}}{N_{ijk}}=R_k \qquad [2]$$

where

R_k represents the overall target sampling rate for the k^{th} ACS second-stage stratum.
$N_{ijk}^{(0)}$ represents the total number of addresses available in the current year's (year 0) ACS first-stage sample for the i^{th} county, j^{th} first-stage ACS sampling stratum, and k^{th} ACS second-stage stratum.
$n_{ijk}^{(0)}$ represents the current year's ACS second-stage sample for the i^{th} county, j^{th} first-stage ACS sampling stratum, and k^{th} ACS second-stage stratum.
N_{ijk} represents the total number of valid addresses on the Main 2007 MAF extract in the i^{th} county, j^{th} first-stage ACS sampling stratum, and k^{th} ACS second-stage stratum.

Within each county, first-stage sampling stratum, and second-stage sampling stratum, an adjustment to R_k is made so that the second-stage POS (r) yields the desired overall sample size as follows:

$$r_{ijk}=\frac{R_k\times N_{ijk}}{N_{ijk}^{(0)}} \qquad [3]$$

Thus, R_k can be written as:

$$R_k=\frac{r_{ijk}\times N_{ijk}^{(0)}}{N_{ijk}} \qquad [4]$$

We have established that we need to find F_{ijk} such that:

$$F_{ijk}\times\frac{N_{ijk}^{(0)}-n_{ijk}^{(0)}}{N_{ijk}}=R_k \qquad [5]$$

Substituting [4] for R_k we get:

$$F_{ijk} \times \frac{N_{ijk}^{(0)} - n_{ijk}^{(0)}}{N_{ijk}} = \frac{r_{ijk} \times N_{ijk}^{(0)}}{N_{ijk}} \qquad [6]$$

We would like to express P_{ijk} in terms of r_{ijk}. Since the probability of not being selected in the ACS second-stage sample is $(1 - r_{ijk})$, we can use the relationship

$$N_{ijk}^{(0)} - n_{ijk}^{(0)} = \left(1 - r_{ijk}\right) \times N_{ijk}^{(0)} \qquad [7]$$

to rewrite [6] as:

$$F_{ijk} \times \frac{\left(1 - r_{ijk}\right) \times N_{ijk}^{(0)}}{N_{ijk}} = \frac{r_{ijk} \times N_{ijk}^{(0)}}{N_{ijk}} \qquad [8]$$

Solving for F_{ijk} we get:

$$F_{ijk} = \frac{r_{ijk}}{\left(1 - r_{ijk}\right)} \qquad [9]$$

Substituting this in [1] we get:

$$P_{ijk} = C \times \frac{r_{ijk}}{\left(1 - r_{ijk}\right)} \qquad [10]$$

We require that

$$\sum_{i=1}^{x} \sum_{j=1}^{y} \sum_{k=1}^{z} P_{ijk} \times \left(N_{ijk}^{(0)} - n_{ijk}^{(0)}\right) = 15{,}000 \qquad [11]$$

Substituting from [10] we get:

$$\sum_{i=1}^{x} \sum_{j=1}^{y} \sum_{k=1}^{z} C \times \frac{r_{ijk}}{\left(1 - r_{ijk}\right)} \times \left(N_{ijk}^{(0)} - n_{ijk}^{(0)}\right) = 15{,}000 \qquad [12]$$

Solving for C we get:

$$C = \frac{15{,}000}{\sum_{i=1}^{x} \sum_{j=1}^{y} \sum_{k=1}^{z} \frac{r_{ijk}}{\left(1 - r_{ijk}\right)} \times \left(N_{ijk}^{(0)} - n_{ijk}^{(0)}\right)} \qquad [13]$$

Therefore the following is the probability of selection for the i^{th} county, j^{th} first-stage stratum, and k^{th} second-stage stratum:

$$P_{ijk} = \frac{15{,}000}{\sum_{i=1}^{x} \sum_{j=1}^{y} \sum_{k=1}^{z} \frac{r_{ijk}}{\left(1 - r_{ijk}\right)} \times \left(N_{ijk}^{(0)} - n_{ijk}^{(0)}\right)} \times \frac{r_{ijk}}{\left(1 - r_{ijk}\right)} \qquad [14]$$

Since F_{ijk} – by construct – when applied to the available universe, $N_{ijk}^{(0)} - n_{ijk}^{(0)}$, yields the same target sampling rate as the ACS rate, R_k, with respect to the overall universe, N_{ijk}, the Original Content Test Base Weight (the inverse of the overall probability of selection) for each record in sub-stratum *k* is calculated as:

$$CTBW_k = \left(C \times R_k\right)^{-1} \qquad [15]$$

In practice, the probability *P(CT)* for each unit was simply calculated as:

$$P(CT) = \mathrm{C} \times \frac{\text{P(in first - stage sample)}}{\text{P(not in second - stage)}}.$$

The Grid/Sequential Questionnaire Test

After the Original Content Test was cancelled, the Grid/Sequential Questionnaire Test was enacted to only evaluate potential response differences from asking the basic demographic items using a grid versus sequential format. The ACS has traditionally used a horizontal grid format, where the names are listed down the side of the page and the questions appear across the top of the questionnaire, to collect the basic demographic data. Census 2010 plans to use a sequential format where each person's data appear in a distinct column, and within each column, the names are at the top and the questions are listed down the page. In support of consistency efforts between the ACS and Census, the ACS tested whether the layout for these items affects response. This test, referred to as the Grid/Sequential Questionnaire Test, was implemented to assess if one format improves data quality and response rates. It was a mail only test and did not include Alaska, Hawaii, or Puerto Rico.

The target universe for the 2007 Grid/Sequential Questionnaire Test consists of all valid, residential housing unit addresses in all county and county equivalents in the United States, excluding Alaska, Hawaii, and Puerto Rico. A sample with a target size of 30,000 was selected from the 2007 ACS Main first-stage sample that was not selected in the ACS second-stage sampling or in any other operation (training, other tests, etc.). Because this test was implemented instead of the Original Content Test, the units selected in that sample were still eligible to be selected for this test. Sample selection used rates that achieve equal overall probabilities of selection within response stratum from the sampling universe in order to have a sufficient amount of data to analyze responses from low response areas as well as the full results. The 2006 ACS Content Test also had a similar sample design. See Asiala (2006) for more details.

The basic experimental design of the test is displayed in Table 2 below.

Table 2: Experimental Design of the Grid/Sequential Questionnaire Test

Grid Questionnaire Design		Sequential Questionnaire Design	
High Response Stratum	Low Response Stratum	High Response Stratum	Low Response Stratum

The sample design requirements for this test were as follows:

- Each of the 15,000 addresses in each treatment sample has an equal overall probability of selection from the sampling universe within response stratum.
- 60% of the sample addresses are selected from the low response stratum and 40% are selected from the high response stratum.
- The sample selection is a paired design. That is, for each selected address, the next address on the list is also selected.
- One address in each pair is randomly assigned to the Grid treatment panel and the other to the Sequential treatment panel.

All census tracts were stratified by their Census 2000 long form mail response rate into either a high or low response stratum. The two strata were defined such that the high response stratum contained housing unit addresses residing in tracts with a mail response rate higher than or equal to 63%. The remaining tracts were assigned to the low response stratum. The 63% cutoff point places approximately 75% of the total number of addresses in the high response stratum.

The Grid/Sequential Questionnaire Test sample was allocated so that the expected number of mail returns was equal within both strata. This will, in expectation, result in similar variances by response stratum.

The sample selection procedure used a paired design in which a sample of 15,000 was selected and the nearest available neighbor was also selected into the Grid Sequential Questionnaire Test sample. Where possible, all pairs of sample records came from the same response stratum/state/county/ACS first-stage sampling stratum/ACS second-sampling stratum combination. Unbiased sampling weights were also calculated. Each sample of 15,000, when weighted, is representative of the universe. The sampling procedure also randomly assigned each sample of 15,000 to one of two treatment panels: grid or sequential questionnaire design.

The sampling frame for the test was the unused portion of the 2007 ACS Main first-stage ACS sample. The sample selection was designed such that, within response stratum, each address in the frame has equal probability of selection from the universe. To achieve this, three stages of sampling needed to be accounted for: 1) ACS first-stage sampling, 2) ACS second-stage sampling, and 3) the Grid/Sequential (or third stage) sample selection. For further details on the sample design, see Joshipura and Hefter (2007b).

Derivation of the Selection Probabilities for the Grid/Sequential Questionnaire Test

The Grid/Sequential Questionnaire Test probability of selection, $P(GS)$, from the records in the ACS first-stage sample not selected in the ACS second-stage sample, within each county/first-stage stratum/second-stage stratum/response stratum can be denoted as: P_{ijkl}.

where

i = 1,...,x (x = # of counties available on the Grid/Sequential Questionnaire sampling frame (*GSFRAME*)).

j = 1,...,y (y = # of substrata in county j)

k = 1,...,z (z = # of second-stage strata in county i and first-stage stratum j)

l = 1, 2 (1 = low response stratum, 2 = high response stratum)

The requirement of each address in the frame having equal probability of selection from the universe leads to the following necessary condition:

$$P_{ijkl} \times \frac{N^{(0)}_{ijkl} - n^{(0)}_{ijkl}}{N_{ijkl}} = \frac{t_l}{N_l} \qquad [1]$$

where

$N^{(0)}_{ijkl}$ represents the total number of addresses available in the current year's (year 0) ACS first-stage sample for the i^{th} county, j^{th} first-stage ACS sampling stratum, k^{th} ACS second-stage stratum, and the l^{th} response stratum.

$n^{(0)}_{ijkl}$ represents the current year's ACS second-stage sample for the i^{th} county, j^{th} first-stage ACS sampling stratum, k^{th} ACS second-stage stratum, and the l^{th} response stratum.

N_{ijkl} represents the total number of valid addresses on the Main 2007 MAF extract in the i^{th} county, j^{th} first-stage ACS sampling stratum, k^{th} ACS second-stage stratum, and the l^{th} response stratum.

t_l represents the sample size for the l^{th} response stratum (i.e. t_1 = 0.6 × 15,000 for the low response stratum and t_2 = 0.4 × 15,000 for the high response stratum)

N_l represents the total number of valid addresses on the Main 2007 MAF extract in each response stratum.

Equation [1] can also be written as:

$$P_{ijkl} \times \frac{\left(1 - r_{ijkl}\right) \times N_{ijkl}^{(0)}}{N_{ijkl}} = \frac{t_l}{N_l} \qquad [2]$$

since

$$N_{ijkl}^{(0)} - n_{ijkl}^{(0)} = \left(1 - r_{ijkl}\right) \times N_{ijkl}^{(0)} \qquad [3]$$

Solving for P_{ijkl} , we get:

$$P_{ijkl} = \frac{t_l}{N_l} \times \frac{N_{ijkl}}{\left(1 - r_{ijkl}\right) \times N_{ijkl}^{(0)}} \qquad [4]$$

Analogous to the calculation of the Original Content Test base weight derived earlier, it then follows from [1] that the Grid/Sequential Questionnaire Test Base Weight (GSBW) for each record in response stratum *l* is calculated as:

$$GSBW_l = \frac{N_l}{t_l} \qquad [5]$$

Therefore, the overall probability from the universe, *P*, was then calculated as follows:

$$P = P(\text{in first-stage sample}) \times P(\text{not in second-stage sample}) \times P(\text{in third stage sample}),$$

where *P(in third-stage sample)* is the probability of selection used to sample addresses from the Grid/Sequential sampling frame $\left(\sum_{ijkl} N_{ijkl}^{(0)} - n_{ijkl}^{(0)}\right)$.

Selected Results for the Grid/Sequential Questionnaire Test

Table 3 below shows the two overall probabilities of selection (POS) from the sampling universe:

Table 3: Overall POS from the Sampling Universe by Response Stratum

Response Stratum	Overall Probability of Selection
High	0.0000604
Low	0.0002842

The total number in sample was 29,998. Both the grid treatment panel and the sequential treatment panel contained 14,999 addresses. There were 9,000 records in the low response stratum and 5,999 records in the high response stratum in each treatment panel. Table 4 shows the percent of the entire sample in each response stratum.

Table 4: Percent in Sample by Response Stratum

Response Stratum	Number in Sample	Percent of Sample
High	11,998	40.0%
Low	18,000	60.0%
Total	29,998	100%

See Joshipura and Hefter (2007c) for further results.

The 2007 ACS Content Test

Once funding was received for fiscal year 2007, we worked quickly to reinstate the experimental testing of the new Field of Degree question in an effort called the 2007 ACS Content Test. This test was designed to assess the effectiveness of obtaining accurate responses to two different versions of a proposed field of degree question on the ACS questionnaire. A secondary goal of this test was to evaluate any change in the tenure distribution caused by changes in question wording that were enacted for the 2010 census and thus, adopted for the 2008 ACS questionnaire. The 2007 ACS Content Test was implemented in addition to the Grid/Sequential Questionnaire Test; both tests were executed in lieu of the Original Content Test.

The 2007 ACS Content Test sample design is largely based on the ACS design in order to simulate the conditions under which the new field of degree question may ultimately be asked. Under this design there is both CATI and CAPI follow-up of non-responding addresses. This test does not include Alaska, Hawaii, and Puerto Rico. This test also includes a telephone re-interview of the topics of interest such as the tenure and Field of Degree questions for all responding households. This test was mailed out in July 2007, so the corresponding ACS panel is to serve as the control for this experiment.

This test has the same target universe, target sample size, and sample design as the Original Content Test. However, the sampling frame is the 2007 Main *and* Supplemental ACS first-stage sample that was not selected in the ACS second-stage sampling or in any other operation (training, other tests, etc.). Distinctions between the two phases were necessary in the probability of selection calculation.

The requirements of the Content Test sample were:

- Each of the 15,000 addresses in each treatment sample is selected with probability proportional to the ACS sample probability of selection.
- The sample selection should be a paired design. That is, for each selected address, the next address on the list is also selected.
- Each treatment panel has the same number of addresses allocated to it.
- Both addresses within a pair are not assigned the same field of degree question.

The sample selection procedure used a paired design in which a sample of 15,000 is selected and the nearest available neighbor is also selected into the Content Test sample. Where possible, all pairs of sample records came from the same state/county/ACS second-sampling stratum combination. Unbiased sampling weights are also calculated. Each sample of 15,000, when weighted, is representative of the universe. Each sample of 15,000 was randomly assigned to one of two treatment panels: version 1 of the field of degree question or version 2.

The Content Test sample design was driven by the requirement that each of the 15,000 addresses in each treatment sample be selected with probability proportional to the ACS sample probability of selection (POS). The frame for this test was the unused portion of the first stage ACS sample. This is the portion of both the Main and Supplemental first-stage sample not selected in the second-stage ACS sample or the 2007 Grid/Sequential Questionnaire Design Test. For further details on sample design, see Joshipura and Hefter (2007d).

Derivation of the Selection Probabilities for the ACS Content Test

Define ACS Content Test POS, *P(CT)* to be:

$$P_{ijkls}\left(CT\right) = C \times F_{ijkls} \qquad [1]$$

where

C represents a constant scalar used to achieve the desired sample size of 15,000

F_{ijkls} represents the factor of P_{ijklr} that is proportional to the ACS second-stage POS for the i^{th} county, j^{th} first-stage ACS sampling stratum, k^{th} second-stage sampling stratum, l^{th} sampling phase, and s^{th} Grid/Sequential response stratum.

i = 1,..., x (x = # of counties available on the content sampling frame).

j = 1,..., y (y = # of first-stage strata county i)

k = 1,..., z (z = # of second-stage strata in county i and first-stage stratum j)

l = 1, 2 (1 = main, 2 = supplemental)
s = 1, 2 (1 = Grid/Sequential low response stratum, 2 = Grid/Sequential high response stratum)

The requirement stated above leads to the following necessary condition:

$$F_{ijkls} \times \frac{N_{ijkls}^{(0)} - n_{ijkls}^{(0)} - m_{ijkls}}{N_{ijkls}} = R_k \tag*{[2]}$$

where

R_k represents the overall target sampling rate for the k^{th} ACS second-stage stratum.

$N_{ijkls}^{(0)}$ represents the total number of addresses available in the current year's (year 0) ACS first-stage sample for the i^{th} county, j^{th} first-stage ACS sampling stratum, k^{th} ACS second-stage stratum, l^{th} sampling phase, and s^{th} response stratum.

$n_{ijkls}^{(0)}$ represents the current year's ACS second-stage sample for the i^{th} county, j^{th} first-stage ACS sampling stratum, k^{th} ACS second-stage stratum, l^{th} sampling phase, and s^{th} response stratum.

m_{ijkls} represents the number in sample in the Grid/Sequential Questionnaire Design Test for the i^{th} county, j^{th} first-stage ACS sampling stratum, k^{th} ACS second-stage stratum, l^{th} sampling phase, and s^{th} response stratum. (Note that this number is 0 when *l* = 2.)

N_{ijkls} represents the total number of valid addresses in the i^{th} county, j^{th} first-stage ACS sampling stratum, k^{th} ACS second-stage stratum, l^{th} sampling phase, and s^{th} response stratum.

As noted earlier, during ACS sampling, within each county, first-stage sampling stratum, and second-stage sampling stratum, an adjustment to R_k is made so that the second-stage POS (*r*) yields the desired overall sample size as follows:

$$r_{ijkls} = \frac{R_k \times N_{ijkls}}{N_{ijkls}^{(0)}} \tag*{[3]}$$

Thus, R_k can be written as:

$$R_k = \frac{r_{ijkls} \times N_{ijkls}^{(0)}}{N_{ijkls}} \tag*{[4]}$$

We have established in [2] that we need to find F_{ijkls} such that:

$$F_{ijkls} \times \frac{N_{ijkls}^{(0)} - n_{ijkls}^{(0)} - m_{ijkls}}{N_{ijkls}} = R_k \tag*{[5]}$$

Substituting [4] for R_k we get:

$$F_{ijkls} \times \frac{N_{ijkls}^{(0)} - n_{ijkls}^{(0)} - m_{ijkls}}{N_{ijkls}} = \frac{r_{ijkls} \times N_{ijkls}^{(0)}}{N_{ijkls}} \tag*{[6]}$$

Since the probability of not being selected in the Grid/Sequential Questionnaire Design Test, within each Grid/Sequential response stratum, is $\left(1 - GPOS_{ijkls}\right)$ we have the following relationship:

$$N_{ijkls}^{(0)} - n_{ijkls}^{(0)} - m_{ijkls} = \left(1 - GPOS_{ijkls}\right) \times \left(N_{ijkls}^{(0)} - n_{ijkls}^{(0)}\right) \qquad [7]$$

where $GPOS_{ijk1s}$ is two times the probability of selection used for the Grid/Sequential Questionnaire Design Test ($GSPOS_{ijk1s}$) since two samples of 15,000 were selected for the Grid/Sequential Test and GPOS $_{ijk2s}$ is equal to zero.

As noted earlier,

$$N_{ijkls}^{(0)} - n_{ijkls}^{(0)} = \left(1 - r_{ijkls}\right) \times N_{ijkls}^{(0)} \qquad [8]$$

We can rewrite [5] as:

$$F_{ijkls} \times \frac{\left(1 - r_{ijkls}\right) \times \left(1 - GPOS_{ijkls}\right) \times N_{ijkls}^{(0)}}{N_{ijkls}} = \frac{r_{ijkls} \times N_{ijkls}^{(0)}}{N_{ijkls}} \qquad [9]$$

Solving for F_{ijkls} we get:

$$F_{ijkls} = \frac{r_{ijkls}}{\left(1 - r_{ijkls}\right) \times \left(1 - GPOS_{ijkls}\right)} \qquad [10]$$

Substituting this in [1] we get:

$$P_{ijkls} = C \times \frac{r_{ijkls}}{\left(1 - r_{ijkls}\right) \times \left(1 - GPOS_{ijkls}\right)} \qquad [11]$$

We require that

$$\sum_{i=1}^{x} \sum_{j=1}^{y} \sum_{k=1}^{z} \sum_{l=1}^{2} \sum_{s=1}^{2} C \times F_{ijkls} \times \left(N_{ijkls}^{(0)} - n_{ijkls}^{(0)} - m_{ijkls}\right) = 15{,}000 \qquad [12]$$

Solving for C in [12], we get

$$C = \frac{15{,}000}{\sum_{i=1}^{x} \sum_{j=1}^{y} \sum_{k=1}^{z} \sum_{l=1}^{2} \sum_{s=1}^{2} F_{ijkls} \times \left(N_{ijkls}^{(0)} - n_{ijkls}^{(0)} - m_{ijkls}\right)} \qquad [13]$$

Let $Q_{ijkls} = \left(N_{ijkls}^{(0)} - n_{ijkls}^{(0)} - m_{ijkls}\right)$

The following probabilities of selection are assigned to each record in the i^{th} county, j^{th} first-stage stratum, k^{th} second-stage stratum, l^{th} sampling phase, and the s^{th} response stratum:

For cases eligible from the Main phase first-stage sample:

$$P_{ijkls} = \frac{15{,}000}{\sum_{i=1}^{x}\sum_{j=1}^{y}\sum_{k=1}^{z}\sum_{l=1}^{2}\sum_{s=1}^{2} F_{ijkls} \times Q_{ijkls}} \times \frac{r_{ijk1s}}{(1-r_{ijk1s})\times(1-GPOS_{ijk1s})} \qquad [14]$$

For cases eligible from only the Supplemental phase first-stage sample:

$$P_{ijkls} = \frac{15{,}000}{\sum_{i=1}^{x}\sum_{j=1}^{y}\sum_{k=1}^{z}\sum_{l=1}^{2}\sum_{s=1}^{2} F_{ijkls} \times Q_{ijkls}} \times \frac{r_{ijk2s}}{(1-r_{ijk2s})} \qquad [15]$$

Note: Using the definition of *Q* in equation [2], we get that $F_{ijkls} \times Q_{ijlks} = R_k \times N_{ijkls} = n_{ijkls}^{(0)}$. Thus the summation $\sum_{i=1}^{x}\sum_{j=1}^{y}\sum_{k=1}^{z}\sum_{l=1}^{2}\sum_{r=1}^{2} F_{ijklr} \times Q_{ijklr}$ can also be written as $\sum_{i=1}^{x}\sum_{j=1}^{y}\sum_{k=1}^{z}\sum_{l=1}^{2}\sum_{r=1}^{2} n_{ijklr}^{(0)}$, which is equal to the total expected sample size for the ACS.

Since F_{ijkls} – by construct – when applied to the available universe, Q_{ijkls}, yields the target ACS sampling rate, R_k, with respect to the overall universe, N_{ijkls}, it follows that the ACS Content Test Base Weight (NCTBW) for each record in sub-stratum *k* is calculated as:

$$NCTBW_k = (C \times R_k)^{-1} \qquad [16]$$

The probability for each unit, *P(CT)*, was therefore calculated as follows:

$$P(CT) = \mathrm{C} \times \frac{\text{P(in first - stage sample)}}{\text{P(not in second - stage)} \times \text{P(not in Grid/Sequential sample)}}.$$

Selected Results for the ACS Content Test

The total number of cases in the ACS Content Test sample is 30,000. Both panels were allocated 15,000 addresses (40% and 60% from the respective response strata). The ACS Content Test was designed such that the percent of the sample in each unique target sampling rate stratum would be similar to the 2007 ACS. The comparison between the two is shown in Table 5 below.

Table 5: Percent of Content Test Sample by Stratum

Second-Stage Sampling Rate	2007 Content Test	2007 ACS
Base Rate	21.59%	21.56%
0.92 × Base Rate	24.59%	24.60%
3 × Base Rate	13.81%	13.82%
1.5 × Base Rate	3.83%	3.82%
0.75 × Base Rate	10.33%	10.33%
0.92 × 0.75 × Base Rate	21.77%	21.78%
10%	4.09%	4.09%

See Joshipura and Hefter (2007e) for further results.

Conclusion

The ACS sample design has provided an excellent jumping off point – or springboard – for these three experimental sample designs. The sampling frames that were derived from it have enabled sample selections that will not overburden the households selected while achieving each design's objectives. Each address selected in these experimental samples is not eligible to be in an ACS sample or an ACS related sample for the next four years. The sampling frames also allow for many sample designs that mimic the ACS at either phase of sampling or uses a constant overall sampling rate for each treatment.

In addition, other surveys may use the ACS to design highly efficient small-scale sample designs in which they can combine the ACS response data. Also, the probabilities of selection already assigned to the first-stage sample easily facilitate these sample designs, whether mirroring the ACS or not.

References

Asiala, M. (2005), "American Community Survey Research Report: Differential Sub-Sampling in the Computer Assisted Personal Interview Sample Selection in Areas of Low Cooperation Rates," Internal Census Bureau Memorandum to R. Singh from D. Hubble, Washington, DC, February 15, 2005.

Asiala M. (2006), "Experimental Design for the 2006 American Community Survey Content Test," *2006 Proceedings of the Joint Statistical Meeting,* American Statistical Association, Washington DC.

Bates, L. M. (2006), "Editing the MAF Extracts and Creating the Unit Frame Universe for the American Community Survey," Internal U.S. Census Bureau Memorandum from D. Kostanich to L. Blummerman, Draft, Washington, DC, September 20, 2006.

Joshipura, M. and Hefter, S. (2006) "Specifications for Selecting the 2007 American Community Survey Content Test Sample –Field of Degree Question Design," 2007 American Community Survey Sampling Memorandum Series #ACS07-S-8a, Internal U.S. Census Bureau Memorandum from D. Whitford to S. Schechter, Draft, Washington, DC, December 27, 2006.

Joshipura, M. and Hefter, S. (2007a) "2007 Content Test Sample Results Documentation Memorandum," 2007 American Community Survey Sampling Memorandum Series #ACS07-S-12, Internal U.S. Census Bureau Memorandum from A. Navarro to D. Whitford, Draft, Washington, DC, January 30, 2007.

Joshipura, M. and Hefter, S. (2007b) "Specifications for Selecting the 2007 Grid/Sequential Questionnaire Test Sample," 2007 American Community Survey Sampling Memorandum Series #ACS07-S-8b, Internal U.S. Census Bureau Memorandum from D. Whitford to S. Schechter, Washington, DC, January 8, 2007.

Joshipura, M. and Hefter, S. (2007c) "2007 Grid/Sequential Questionnaire Test Sample Results Documentation Memorandum," 2007 American Community Survey Sampling Memorandum Series #ACS07-S-13, Internal U.S. Census Bureau Memorandum from A. Navarro to D. Whitford, Washington, DC, February 16, 2007.

Joshipura, M. and Hefter, S. (2007d) "Specifications for Selecting the 2007 American Community Survey Content Test Sample," 2007 American Community Survey Sampling Memorandum Series #ACS07-S-14, Internal U.S. Census Bureau Memorandum from D. Whitford to S. Schechter, Draft, Washington, DC, June 29, 2007.

Joshipura, M. and Hefter, S. (2007e) "2007 American Community Survey Content Test Sample Results," 2007 American Community Survey Sampling Memorandum Series #ACS07-S-15, Internal U.S. Census Bureau Memorandum from A. Navarro to D. Whitford, Washington, DC, June 8, 2007.

U.S. Census Bureau (2006), "Design and Methodology: American Community Survey", U.S. Government Printing Office, Washington, DC, 2006 (http://www.census.gov/acs/www/Downloads/tp67.pdf).

Zimolzak, M. and Rose, S. (2007), "Customer Requirements Documents for American Community Survey Data Products", Version 1.0, Internal Census Bureau Memorandum, Washington, DC, June 13, 2007.