Experimental Design for the 2006 American Community Survey Content Test

Mark E. Asiala, Alfredo Navarro
US Census Bureau, Room #4H054E
4600 Silver Hill Rd, Washington, DC 20233
mark.e.asiala@census.gov

Abstract

The 2006 American Community Survey (ACS) Content Test is a large household sample survey (over 60,000) designed to test proposed changes to the ACS questionnaire for 2008. The original base sample design is a stratified primary sample unit design that builds on the sample design of an earlier survey, the Census 2000 Supplementary Survey. In response to changing field requirements, several changes were implemented to address total workload size and other operational issues. Additional stratification and sampling for the personal follow-up phase of the survey was added late in the design phase to help reduce the total workload. A balanced split-panel assignment was designed to help minimize the number of field representatives (FR's) necessary for the survey in the less densely populated areas avoiding duplication of geographic assignment areas for experimental and control panel FR's.

Key Words: content test, experimental design, split-panel

1 Background

The American Community Survey is part of the Census Bureau's plans for a re-engineered 2010 Census. The ACS will collect long-form (sample) data on an annual basis in order to produce single and multi-year estimates which are comparable to the long-form estimates traditionally produced after each decennial census.

In preparation for the 2008 ACS, the Census Bureau began a critical review of all questions included on the ACS. The review included looking at several different sources of information that could speak to the quality of the data for indivudual questions or topics contained on the ACS. Subject matter specialists within the Census Bureau as well as from other Federal agencies participated in the review through an Office of Management and Budget (OMB) Interagency Committee on the ACS. As a result of that review several questions were identified for testing in the 2006 ACS Content Test.

The 2006 ACS Content Test had three high-level objectives:

1. Per specific content areas, can changes to the response categories, question wording and redefinition of underlying constructs improve the quality of the collected data?

2. Do changes in layout of the mail form necessary to accommodate the modified basic demographic questions impact response at a unit or item level?

3. If the ACS adopts the modified content, thus increasing the overall length of the questionnaire, can the ACS contain mailing costs by dropping one piece of the mailing package (the questionnaire instruction booklet) without adversely impacting data quality?

Each objective needed to be tested which suggested a split-panel experimental design. Each high-level objective translated into a dimension forming a 2 x 2 x 2 design.

2 Methodology

The use of split-panel design for content tests at the Census Bureau is well established. In order to work out the details of the design, the high-level objectives needed translating into specific functional requirements.

2.1 Design Requirements for the Test

The original design requirements for the test came from a combination of the functional requirements dictated by the high-level objectives and from the realities of trying to conduct the test in a cost efficient manner.

The sample size for the content test was driven by the unemployment content because of the importance of this area. Census staff also were interested

This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical issues are those of the authors and not necessarily those of the U.S. Census Bureau.

if there was a differential effect for the content and form changes by response rate stratum (high/low response). The remaining requirements defined certain universe restrictions and field requirements.

1. The experimental design must be able to detect a 0.5 percentage point difference in the estimate of unemployment at a national level (at a 10 percent significance level).

2. The sample must be stratified by high/low mail response. The expected number of mail returns from the two response strata must be equal at the national level. The sample will be assigned to a response strata at the tract level.

3. Certain large counties should be included with certainty because of the concentration of certain demographic groups of interest.

4. Alaska, Hawaii, and Puerto Rico are not in the scope of the content test and should be excluded from the universe.

5. The estimated person follow-up assignment workload for the experimental panel must be 10–15 cases or a multiple thereof to maximize field staff efficiency. (this threshold was later modified)

2.2 Experimental Design

The development of the experimental design followed past content tests at the bureau and other surveys by planning a split-panel design built on clustering counties into primary sampling units (PSU). The base PSU design works well for achieving national estimates and for clustering workloads of the non-response follow-up.

To help expedite the design process, the PSUs used for the content test were the same PSUs originally formed for the Census 2000 Supplementary Survey (C2SS) (Shoemaker, 1998). The C2SS was a national state-based design to test the feasibility of employing ACS data collection methods at a national level and to make certain comparisons between the data collected from the C2SS and the Census 2000 sample data. For these reasons, it seemed to be a natural fit for our needs. The PSU strata for the content test were also borrowed from the original strata of the C2SS and were amended as necessary.

2.2.1 Requirement #1

Initial calculations using simple random sample variance calculations indicated that a split-panel design based on 50,000 sample addresses split across the two content panels would meet the detectability criterion (Requirement #1). Later, the final design would be re-evaluated using a better stratified variance estimator to ensure that the design variance estimate would satisfy the detectability criterion.

2.2.2 Requirement #2

To meet the high/low response strata requirement, we classified all tracts into a stratum based on their Census 2000 long-form response rate, defined as the number of interviews over the eligible respondents (removing unmailable addresses and vacant housing units). After sorting all tracts by their response rate in ascending order, the cutoff for the high/low response stratum was drawn where 25 percent of the total number of housing units were located in the low response stratum. Conversely, 75 percent of the total number of housing units reside in tracts in the high response stratum. These same high/low response strata were used in the 2003 ACS Voluntary Test (US Census Bureau, 2003).

Once all tracts had been classified, an overall mail return rate (defined as the number of mail returns over the size of the entire sample) was created for each stratum using a combination of historical ACS mail return rates and modeled rates based on Census 2000 long-form return rates. The modeled rates were produced by creating a simple linear model between the ACS mail return rates and the long-form return rates at the tract level for tracts which had sample in both. The model was then applied to the tracts where there was not any historical ACS rates. Of the approximately 65,000 tracts in the US, only 15,000 of the tracts required use of the modeled rates.

Using the overall mail return rate by stratum and the current number of addresses by stratum plus the total sample size, the sampling rates for each response stratum were derived in order to achieve the goal in Requirement #2 of equal expected mail returns by stratum.

2.2.3 Requirement #3 and #4

Little needed to be done for these two requirements. The large counties that were required were ones in large metropolitan areas with varying demographics for Hispanic, foreign born, and other characteristics. All proposed PSU designs were evaluated for this requirement. Requirement #4 was simply an exclusion of Alaska, Hawaii, and Puerto Rico as a cost saving measure.

2.2.4 Requirement #5

Knowing the sampling rates as the result of the work on #2 allowed us to procede with clustering the PSUs into strata. Using the sample sizes and mail return rate data calculated, we were able to further estimate the non-response workloads for both telephone and personal visit by PSU.

Selfrepresenting PSUs were identified as those PSUs which were selfrepresenting in the C2SS sample design and whose expected personal visit workload met the requirement of having 10–15 addresses or a multiple thereof in the content test. This yielded 185 selfrepresenting PSUs. All remaining PSUs where then defined to be non-selfrepresenting and were clustered with other non-selfrepresenting PSUs into estimation strata. These non-selfrepresenting strata were formed by building from the strata defined for the C2SS design and then collapsing as necessary until the estimated workload of the strata was at least 20–30 addresses. The sample design would select two PSUs per stratum so in expectation each PSU would satisfy the workload requirement of having 10–15 non-response addresses. It was acknowledged, however, that the actual workloads could be greater or smaller than the target value.

From each non-selfrepresenting stratum, two PSUs were selected using a probability proportional to size sampling method. The measure of size for each PSU was the 2003 Intercensal Population Estimates which are produced annually by the Population Estimate Program of the Census Bureau.

Within each PSU, the original design called for a fully interpenetrated design where pairs of nearby addresses would be selected with one address being assigned to the control content and one address being assigned to the experimental content.

The design arrived at through this process was then checked to see if it met the detectibility requirement using the more accurate stratified estimator described in Section #3. Once this was verified, the design and its associated non-response workloads was presented to the field staff for their buy-in.

2.3 Amendments to the original design

Review of the field staff workloads led to three revisions in the original design. The first revision was introduced because the field headquarters staff responded to the initial design as having too high of a non-response workload. In order to accomodate this feedback and to control costs, a design change was introduced to conduct the personal visit or Computer Assisted Personal Interviewing (CAPI) non-response work in only a sampled subset of the sample PSUs. To minimize the risk that sub-sampling PSUs would significantly affect our ability to analyze data specific to the foreign-born population, the PSUs were stratified by the estimate of foreign born within the PSU.

First, all sample PSUs were sorted in a descending sort by the estimated number of foreign born within the PSU based on Census 2000 sample data. The top 20 PSUs in this sort were then selected with certainty to have CAPI follow-up at a 1-in-2 subsampling rate. These top 20 PSUs were collectively called size stratum one. Size stratum two was comprised of the remaining 393 PSUs. The PSUs in size stratum two were sorted by state and foreign born. One out of every three PSUs was then selected to have CAPI follow-up out of the second size stratum. The sub-sampling done for the CAPI follow-up in these PSUs was set at 2-in-3 to help offset the sampling of the PSUs for CAPI. No CAPI work was done in those PSUs not selected for follow-up. Because of the interpenetrated design, however, it was decided that the design objectives could still be met.

The second revision was introduced because of staffing concerns raised by the overlap in timing of the content test non-response follow-up and the Annual Social and Economic Supplement (ASEC) of the Current Population Survey.

The decision was made to alter the normal three-month data collection period (mail, telephone follow-up, personal visit follow-up) for the content test to a two-month data collection period by removing the telephone follow-up stage. Shortening data collection by one month eliminated the overlap with ASEC. Specifically dropping telephone follow-up allowed the test to maintain the entire mail non-response universe as opposed to only including non-respondents for whom we had a telphone number. In order to address the loss of sample interviews as the result of eliminating the telephone follow-up, the inital sample was inflated so that the expected number of mail returns would make up the difference. This raised the total initial sample from 50,000 to 63,000. To avoid a similar increase in the personal visit workload, the non-response universe was further subsampled in order to cap the workload estimates.

The third revision was introduced for a small number of PSUs where the expected workload was small (fewer than 10 addresses) per content panel. One of the requirements for the field staffing was that a single interviewer could not work on both the control and experimental content. For a small number of PSUs, this would have meant training interviewers for the experimental content follow-up who had an expected workload of only a few addresses. Thus to improve efficiency, the decision was either to allow interviewers to work on both panels or to cluster the sample in these small PSUs to be either control or experimental content rather than be fully interpenetrated.

It was decided to go with the clustering approach to avoid the contamination that could result from allowing interviewers to work on both content panels. There were a total of 60 small PSUs that needed to be divided between the control and experimental content. The assignment was conducted using a linear programming method in order to minimize the difference between the two groups on a set of variables correlated to variables of interest in the content analysis and also for the expected sample sizes, nonresponse workloads, and the number of interviews. That work is described in more detail in Section #4.

With the final design created, it was re-evaluated on the detectability criterion using a stratified variance estimator. That work is described in the next section.

3 Variances

The design of the 2006 ACS Content Test must meet the functional requirement of being able to detect a 0.5 percent difference between the control and test panels for the unemployment question. This was originally characterized as being a 5 percent characteristic but was later revised using more recent data as approximately a 7.6 percent characteristic. In order to ensure that the design would meet this functional requirement, a stratified variance estimator extended from Tersine and Starsinic (2003) was applied. That variance estimator used for the production ACS was adapted to account for the particulars of the content test design.

The variance estimator in Tersine and Starsinic accounts for four possible strata: mail returns, Computer Assisted Telephone Interviewing (CATI) interviews, mailable CAPI interviews and unmailable CAPI interviews. The estimator also accounts for occupancy rates in its estimate of the variance of a hypothetical population characteristic.

We extend this estimator by including the mail response strata and the PSU CAPI sampling strata that is a part of the content test design. The strata for CATI and for unmailables can be eliminated since we will not be subsampling CATI and the unmailable CAPI subsampling rate will be equal to the mailable CAPI subsampling rate. Thus the mail and CATI components can be combined into one strata and the CAPI component can be similarly combined. (Note: in the final design the CATI component was completely dropped.)

3.1 Derivation of the Variance Estimator

The variance estimator is for a hypothetical population characteristic, P. The estimate of the overall proportion, $\hat{P}$, is a weighted estimate of the individual proportions from each strata based on the percent of the total occupied housing units that fall within each strata.

To derive the variance estimator, first we define our strata:

Response Strata: Let $j = 1, 2$ designate the high or low response stratum respectively. This captures the difference in weights due to the different sampling rates for the two response strata.

Size Strata: Let h designate the Top-20 Size Strata. PSUs in the top 20 PSUs when sorted by foreign born have $h = 1$ and all other PSUs have $h = 2$. Let n_h denote the number of PSUs in each stratum and $i = 1, \ldots, n_h$ denote the individual PSUs in each stratum. This strata captures the difference in weights due to the sampling of the PSUs selected for CAPI follow-up.

CAPI Strata: The CAPI strata are designated in the variables by a nc marking for Non-CAPI and c for CAPI. This captures the difference in weights due to the CAPI sub-sampling. Since the CAPI sub-sampling rate differed by Size Stratum, this strata is dependent on the Size Strata.

To define the weighted estimator and the variance estimator we define the following variables. Where possible, we try to maintain the notation used by Tersine and Starsinic.

f_j Sampling fraction for response stratum j.

f_h Sampling fraction for PSUs selected for CAPI in size stratum h.

$f_p^{(h)}$ Sampling fraction for CAPI (p) in size stratum h.

$n_{hij}^{(mt)}$ Number of occupied sample unit representing the non-CAPI component in size stratum h, PSU i, and response stratum j.

$n_{hij}^{(p)}$ Number of occupied sample unit representing the CAPI component in size stratum h, PSU i, and response stratum j.

R_{po} Proportion of occupied CAPI cases interviewed (assumes all vacant CAPI cases interviewed).

δ_{hi} Binary indicator designating whether PSU i in size stratum h is selected for CAPI follow-up.

N_{per} Number of persons in the proportion universe per occupied housing unit.

For a given size stratum h, PSU i, and response stratum j, the weighted estimate of the number of persons in the labor force, N_{hij} for that combination hij is calculated by multiplying the weighted estimate of the number of occupied housing units by the average number of persons in the labor force per occupied housing unit, N_{per}. In terms of the defined variables, N_{hij} is defined as:

$$N_{hij} = N_{per}\left[n_{hij}^{(mt)} + n_{hij}^{(p)} R_{po}\left(\frac{\delta_{hi}}{f_h f_p^{(h)}}\right)\right]$$

Note that in N_{hij}, we have summed across the CAPI strata to include both the non-CAPI stratum (interviews from mail and telephone) and the CAPI stratum. The underlying assumption is that all vacants are determined in the CAPI phase of data collection. Thus the non-CAPI stratum count of interviewed housing units, $n_{hij}^{(mt)}$, assumes that the total interviewed housing units are all occupied but the CAPI stratum count of interviewed housing units is multiplied by the factor of percent occupied in the CAPI stratum, R_{po}, to derive the count of occupied housing units.

The weighted total for the number of people who are employed for a given stratum (size, PSU, and response) is calculated by multiplying the stratum counts by the non-CAPI characteristic proportion, $\hat{P}_{hij}^{(mt)}$ and the CAPI characteristic proportion, $\hat{P}_{hij}^{(p)}$. The estimate for the number of employed persons is then

$$A_{hij} = N_{per}\left[n_{hij}^{(mt)}\hat{P}_{hij}^{(mt)} + n_{hij}^{(p)} R_{po}\left(\frac{\delta_{hi}}{f_h f_p^{(h)}}\right)\hat{P}_{hij}^{(p)}\right]$$

The estimate for the unemployment rate, defined as the estimate of the number of people employed divided by the estimated labor force, is the basic ratio:

$$\hat{P} = \frac{\sum_{h=1}^{2}\sum_{i=1}^{n_h}\sum_{j=1}^{2} A_{hij}}{\sum_{h=1}^{2}\sum_{i=1}^{n_h}\sum_{j=1}^{2} N_{hij}}$$

Assuming independence among the strata, the variance of $\hat{P}$ is simply the sum of the strata variances. It is further assumed that $\hat{P}_{hij}^{(mt)} = \hat{P}_{hij}^{(p)} = P$ where P is the population parameter for all h, i, and j. Thus the variance of the strata component of $\hat{P}$ is PQ/n where n is the stratum sample size.

The variance formula thus becomes,

$$\text{Var}\left(\hat{P}\right) = \frac{PQ}{N_{per}}\times \frac{\sum_h\sum_i\sum_j\left(\frac{1}{f_j}\right)^2\left[n_{hij}^{(mt)} + n_{hij}^{(p)}R_{po}\left(\frac{\delta_{hi}}{f_h f_p^{(h)}}\right)^2\right]}{\left[\sum_h\sum_i\sum_j\left(\frac{1}{f_j}\right)\left[n_{hij}^{(mt)} + n_{hij}^{(p)}R_{po}\left(\frac{\delta_{hi}}{f_h f_p^{(h)}}\right)\right]\right]^2}$$

This formula was then used to evaluate alternatives considered along the way to the final design.

3.2 Application

The application of the variance formula involves the use of several parameters, some of which are fixed and some of which can be varied for different design options.

The sampling fraction by response stata use the base sampling rates $f_j = 0.0001016839$ for $j = 1$ (high) and 0.0005151840 for $j = 2$ (low).

$f_h = 1, 2, 3, 5$, or 4 depending on the option.

$f_p^{(h)} = 1, 2/3, 1/2$, or $1/3$ depending on the option.

$n_{hij}^{(mt)}$ = projected number of mail interviews plus the CATI interviews assuming that all responses in the mail and CATI come from occupied housing units.

$n_{hij}^{(p)} = nP_o - n_{hij}^{(mt)}$ = assumes 90 percent of total sample n is occupied and subtract off the non-CAPI interviews to get the estimate of occupied CAPI sample. Multiplying this by R_{po} gets the number of occupied CAPI interviews.

$\delta_{hi} = 0$ or 1 depending on the PSU and the option.

$P = 0.076$, the ACS 2003 estimate of percent of Civilian labor force which was unemployed.

$N_{per} = 1.33$ the average number of persons in the Civilian labor force per occupied housing unit.

The final option has a PSU follow-up sampling fraction, f_h, equal to 1 and $1/3$ for h equal to 1 and 2 respectively. The within PSU CAPI sub-sampling fraction for size strata h was equal $1/2$ and $1/3$ respectively. The use of this estimator allowed us to compare the relative variances of different options that were considered for the sub-sample of PSUs selected for follow-up and different sampling rates.

4 Balancing Problem

The last requirement that was imposed on the design was to accomodate the request from field staff to designate PSUs whose expected non-response follow-up in the CAPI phase of data collection was less than 10 addresses to be either solely control content or experimental content. This constraint was given to alleviate the problem of needing to train separate field staff for each treatment in these small PSUs where it was expected that an interviewer would have only a partial workload. By designating these PSUs as being non-interpenetrated, it would be more cost effective to have the interviewer perform more interviews after incurring the fixed cost of the training. It was also expected that this would improve the quality of the data collection as the interviewer performed a greater number of interviews and became familiar with the changes to the content.

This new criteria led to the identification of 60 PSUs which needed to be split evenly between the control and experimental content panels. The issue becomes what does “evenly” mean besides having 30 PSUs in one panel and 30 in the other? We identified early on that we wanted to have equal sample sizes and equal expected number of interviews between the two groups. Collectively, we refer to these sample sizes, expected interviews, and workloads as our sample design parameters. Some further exploration of the problem led to an adaption of a technique used in the 1986 Census Test. In the design of that test, a set of “disturbing variables” of interest were identified and the design tried to equalize the panels in relation to these variables (Navarro, 1984).

The variables of interest for our work mirrored the tested content. After some debate, we identified the following variables:

1. Percent of labor force which are unemployed
2. Percent of population which are foreign born
3. Percent of population which are high-school graduates
4. Percent of population which are disabled
5. Percent of population which are in poverty
6. Percent of housing units which are owner occupied
7. Percent of population which are Hispanic

For all estimates, the Census 2000 long-form data was used. Fully stated, our set of goals for dividing the PSUs into the two content panels was as follows:

1. Each PSU must be assigned to exactly one panel.
2. The two groups should contain an equal number of PSUs if possible.
3. The two groups should have similar sample design characteristics. This includes similar samples sizes, similar projected mail interviews and similar CAPI workloads.
4. The variance of the sample sizes, mail interviews, and workloads should be similar for the two groups.
5. The simple mean of the 7 variables identified should be approximately equal. This should be done giving a relative priority to the importance of the variables if needed to ensure the best fit for the variables which have the highest priority.
6. The variance of the values of each of the 7 disturbing variables within each group should be approximately equal to the corresponding variance in the other group.

The problem lended itself well to a linear programming solution with each goal contributing to a constraint.

The problem can be characterized in a classic manner. We first define a few variables:

x_i Binary indicator for PSU i assigned to control panel (1=Control 0=Experimental)

s_{gi} Value of sample design parameter g for PSU i. The sample design parameters are Sample Size, Mail Interviews, CAPI workload.

v_{ik} Value of the k-th variable for PSU i

M_k^x Mean of variable k over control panel

V_k^x Variance of variable k over control panel

M_k^y Mean of variable k over experimental panel

V_k^y Variance of variable k over experimental panel

M_k Population Mean of variable k

V_k Population Variance of variable k

n The total number of PSUs (range of i)

m The number of variables (range of k)

$w^M_{v,k}$ The relative weight to consider for the mean of variable k

$w^M_{s,g}$ The relative weight to consider for the mean of sample design parameter g

$w^V_{v,k}$ The relative weight to consider for the variance of variable k

$w^V_{s,g}$ The relative weight to consider for the variance of sample design parameter g

The solution to the problem will be the assignment of a value of 1 or 0 to each of the x_i. By design we meet the criteria that each PSU is assigned to exactly one group. We introduce one constraint to the solution to achieve the second criteria of an equal number of PSUs per group.

$$\sum_{i=1}^{n} x_i = \frac{n}{2}$$

The other criteria are met by defining the appropriate objective function which is to be minimized. Given that we were introducing several requirements, a relative weight was assigned to each goal which could be tweaked if necessary in order to reflect the relative importance of each goal.

The tests for equal means for the variables and sample design parameters is equivalent to testing that the sum of the parameter over all PSUs in the one group is equal to one-half of the total sum across both groups. This yields the following objective functions:

$$\begin{aligned} \sum_{i=1}^{n} (2x_i - 1)\, s_{gi} + N_{s,g} - P_{s,g} &= 0 \\ N_{s,g}, P_{s,g} &\geq 0 \\ \mathrm{Min}_x \,(N_{s,g} + P_{s,g}) & \end{aligned}$$

The variables $N_{s,g}$ and $P_{s,g}$ represent the absolute negative or positive variance from achieving equality in the first equation. In the final solution, both values cannot be non-zero. In order to optimize over the three (possibly) competing sample design parameters, we minimize over the weighted sum of the $N_{s,g}$ and $P_{s,g}$ using the relative weights $w_{s,g}$.

$$\mathrm{Min}_x \left\{ \sum_{g=1}^{3} (N^M_{s,g} + P^M_{s,g}) w_{s,g} \right\}$$

Similarly we have the following objective function for the disturbing variables:

$$\begin{aligned} \frac{2\sum_{i=1}^{n} x_i v_{ik}}{\sum_{i=1}^{n} v_{ik}} + N^M_{v,k} - P^M_{v,k} &= 1 \\ N^M_{v,k}, P^M_{v,k} &\geq 0 \\ \mathrm{Min}_x \left(N^M_{v,k} + P^M_{v,k}\right) & \end{aligned}$$

This minimization is taken across all variables using a weighted sum like what was done for the sample design parameters.

$$\mathrm{Min}_x \left\{ \sum_{k=1}^{M} \left(N^M_{v,k} + P^M_{v,k}\right) w_{v,k} \right\}$$

Initial optimizations were performed investigating just equalizing the means indicated for the two groups. It was unknown how well the optimization may work. After these inital runs which achieved very good optimization on the two groups for equalizing means a new requirement was added that tried to equalize the variance for both the sample design parameters and the variables within the groups.

At first look, linear programming appears to be ill suited for trying to equalize variances because variances involve the sum of quadratic terms. The problem is simplified, however, by treating the set of 60 PSUs as your population and any group of 30 PSUs drawn from it as your sample. A well drawn group of 30 PSUs should then have a variance which is similar to the population variance of the 60 PSUs.

For the sample design parameters, we define the population mean for the sample design parameter g as $M_{s,g}$ and the population variance as $V_{s,g}$. Our goal then is to achieve the following:

$$\begin{aligned} \frac{(2/n)\sum_{i=1}^{n} x_i (s_{gi} - M_{s,g})^2}{V_{s,g}} + N^V_{s,g} - P^V_{s,g} &= 1 \\ N^V_{s,g}, P^V_{s,g} &\geq 0 \\ \mathrm{Min}_x (N^V_{s,g} + P^V_{s,g}) & \end{aligned}$$

The first fraction is simply the variance of the control group for a known population mean and our goal is to achieve the ratio of the variance of the control group to equal the population variance. As a consequence of this, we will also achieve the variance of the experimental group to be equal to the population variance. As can be seen, this can be optimized using a linear programming optimization. Minimizing over all g is obtained by minimizing the weighted sum:

$$\mathrm{Min}_x \left\{ (N^V_{s,g} + P^V_{s,g}) w_{s,g} \right\}$$

Using an analogous derivation to that for the variance in the sample design properties, the goal programming problem is defined for variable k as

$$\frac{(2/n)\sum x_i(v_{i,k} - M_{v,k})^2}{V_{v,k}} + N^V_{v,k} - P^V_{v,k} = 1$$
$$N^V_{v,k}, P^V_{v,k} \geq 0$$
$$\text{Min}_x(N^V_{v,k} + P^V_{v,k})$$

Finally, combining the problems for all variables $k = 1, \dots, m$ using the weights $w_{v,k}$

$$\text{Min}_x \left\{ \sum_{k=1}^{m} (N^V_{v,k} + P^V_{v,k}) w_{v,k} \right\}$$

The requirements to equalize the means was then combined with the requirements of equalizing the variances. Through a series of trial and error optimizations using different relative weights, the final solution was obtained using the weights

$$\begin{aligned} w^M_{s,g} &= 4 \qquad g = 1, \dots, 3 \\ w^V_{s,g} &= 1 \qquad g = 1, \dots, 3 \\ w^M_{v,k} &= 4 \qquad k = 1, \dots, 7 \\ w^V_{v,k} &= 1 \qquad k = 1, \dots, 7 \end{aligned}$$

This solution provides for equal weighting of the individual sample design charactertics and the variables. It does, however, place a higher relative weight on equalizing the means over the variances. This combination produced the best set of comparisons between the two groups considering that the means were the most important to our analysis.

5 Conclusion

The design of the 2006 ACS Content Test posed several challenges after the creation of the initial design. By conducting the personal visit CAPI interviewing in only a sub-sample of the full PSUs selected for sample, we were able to reduce costs while still maintaining a clustered sample for cost efficiency. The expansion on the variance estimator presented in Tersine and Starsinic allowed us to evaluate each option and its impact on the variances. At the last stage, the use of linear programming allowed us to construct two very balanced groups for those small PSUs where we expected a small CAPI workload.

There were some noted limitations to the variance estimator as the design progressed, however. As our non-response follow-up became more clustered, it became more difficult to account for the design effect of that clustering. Some of the parameters we took as constants could have been more geographically tailored in order to produce a more accurate estimate. In the end, however, it was decided that the variance estimate was giving us the accuracy needed in order to evaluate one option against another.

The linear programming solution proved to be an excellent method to balance the non-interpenetrated PSUs. The work done on this portion of the design will have immediate implications for any work done on future content tests with similar requirements for interpenetration.

References

U.S. Census Bureau. (2003). "Meeting 21st Century Demographic Data Needs—Implementing the American Community Survey Report 3: Testing the Use of Voluntary Methods". U.S. Census Bureau, Washington DC

Navarro, A. (1984). "Jersey City Split-Panel - Pilot Study". *SMD 1985 PRETEST MEMORANDA SERIES #K-4.* Washington DC: U.S. Census Bureau. Internal Memorandum.

Tersine, A. & Starsinic M. (2003). "Optimum Non-Response Subsampling Rate for the American Community Survey". *2003 Proceedings of the Joint Statistical Meetings.* Washington DC: American Statistical Association.

Shoemaker, H. (1998). "Documentation of PSU Definitions, Stratification, and Selection for the ACS National Sample 2000–2002". *ACS MEMORANDA SERIES #S-16.* Washington DC: U.S. Census Bureau. Internal Memorandum.