Allocation of Sample for the 2010 Redesign of the Consumer Expenditure Survey October 2012

Stephen Ash[1], Brian Dumbacher[1]
David Swanson[2], Barry Steinberg[2], Sally Reyes-Morales[2]
[1]U.S. Census Bureau, 4600 Silverhill Road, Washington D.C. 20233
[2]Bureau of Labor Statistics, 2 Massachusetts Avenue, NE, Washington D.C. 20212

Abstract
This paper discusses the allocation of sample of the Consumer Expenditure Survey (CE) for the 2010 redesign. CE is a national survey of U.S. households that measures the expenditures made by households on different categories of items. The survey actually consists of two surveys. Small expenditures are collected in the CE Diary Survey and larger expenditures are collected in the CE Quarterly Interview Survey. Both CE surveys use a two-stage sample design, where the first-stage is a sample of counties or groups of counties and the second-stage is a sample of households within the counties or groups of counties.

In this paper we discuss methods for optimally allocating the survey's overall sample to the two surveys where optimality is with respect to minimizing the survey's overall variance. Then, we compare several alternative methods for optimally allocating the sample to the strata of the first-stage sample design.

Key Words: Neyman allocation, Proportional allocation.

1. Introduction

1.1 Background on the Consumer Expenditure Survey
The Consumer Expenditure Survey (CE) is a nationwide household survey conducted jointly by the Bureau of Labor Statistics and the U.S. Census Bureau that collects information on what Americans purchase. The CE survey is composed of two parts, the Diary Survey (Diary) and the Quarterly Interview Survey (Interview). Diary mainly collects information about small frequently purchased items such as food and clothing, while Interview mainly collects information about big-ticket items such as property, automobiles, and regular expenses such as rent and utility bills. The results of the two surveys are combined to produce the estimates of total expenditures.

The universe of interest is the civilian non-institutional population of the U.S., and the survey's sampling frame is a list of addresses generated from the 2010 Decennial Census files with updates from the United States Postal Service Delivery Sequence File and other sources. The unit of interest is the consumer unit, which is defined as the set of people who share living expenses. For convenience and ease of understanding, we will use the term household in place of consumer unit for the remainder of the paper.

CE uses a two-stage sample design. In the first stage, the population is divided into Primary Selection Units (PSUs), which are counties or groups of counties. The PSUs are

stratified by geography and characteristics related to total expenditures. Then a sample of PSUs is selected with probability proportional to size. Some PSUs are so large that they are considered their own strata and have probability of selection equal to one. We refer to these PSUs as "self-representing" PSUs. The remaining PSUs are selected one from each sampling stratum and have a probability of selection less than one. We refer to these PSUs as "non-certainty" or "non-self-representing" PSUs. Both Interview and Diary use the same first-stage sample design.

For the second-stage sample design, addresses are selected within each PSU with systematic random sampling from an ordered list (*sys*). The list of addresses associated with households is sorted by variables that include geography, family size, owner/renter status, property value, and monthly rent.

1.2 First Allocation Problem: Between the Surveys

The first allocation problem is: How should CE allocate its sample between Interview and Diary? Table 1 shows how the sample has been allocated in the past.

Table 1: Historical Allocations

Calendar Year	Households		Interview-to-Diary Ratio
	Interview	Diary	
2005 2	9,804	15,126	1.97
2006 2	8,867	14,455	2.00
2007 2	7,335	13,747	1.99
2008 2	7,545	14,179	1.94
2009 2	8,029	14,623	1.92
Average 2	8,316	14,426	1.96

The target Interview-to-Diary ratio of 2-to-1 was determined nearly thirty years ago, and in 1999 a team at the Bureau of Labor Statistics (Swanson 1999) investigated whether other allocations would produce smaller standard errors of CE estimates. They found that the standard error of CE's most important estimate – the average annualized expenditure per household nationwide – would have been minimized by increasing the ratio to 2.80, but the improvement would have been very small. All in all, there was no compelling reason to change the allocation ratio.

The research of this paper (2012) recommends using the current 2-to-1 Interview-to-Diary allocation ratio. Increasing the ratio to something in the range 2.12 to 2.29 would minimize the standard error, but the gain would be very small. According to our model, the standard error is very insensitive to changes in sample allocation. As a function of the allocation ratio, the standard error is relatively flat at the minimal value. This finding explains why previous research found different allocation ratios as the optimal allocation.

1.3 Second Allocation Problem: Between the First-Stage Units

Historically, the Interview sample size has been allocated proportional to the size of the first-stage strata. The allocation was the same for Interview and Diary. For the current research, we wanted to consider other methods for allocating the sample in order to make the sample design more efficient.

We found that a variation of Neyman allocation for a two-stage sample design did marginally better than the prior method of proportional allocation to the strata.

2. Allocation to the Surveys

2.1 Describing the Optimization Problem

Let $\bar{y}$ denote the average annualized expenditure per household nationwide. To analyze how a change in sample allocation would affect this statistic's variance, we express the standard error of $\bar{y}$ as a function of the annual number of Interview and Diary households, n_I and n_D. Similar to the research of Swanson (1999), we model the standard error by

$$SE(\bar{y}) = \sqrt{\frac{\sigma_I^2}{n_I} + \frac{\sigma_D^2}{n_D} + 532^2},$$

where σ_I^2 / n_I and σ_D^2 / n_D represent the second-stage variances for Interview and Diary, respectively. The sample sizes n_I and n_D are assumed to vary, but the variances σ_I^2 and σ_D^2 are fixed. The value 532^2 represents the first-stage variance for both surveys. We regard the first-stage variance as fixed because changing the Interview-to-Diary allocation ratio does not change the first-stage sample design.

Because the annual CE budget is fixed, increasing the sample size for Interview necessarily means decreasing the sample size for Diary, and vice versa. We incorporate this constraint into our model for calendar year Y by requiring the numbers of households to satisfy

$$c_I n_I + c_D n_D = B_Y,$$

where c_I and c_D are the costs of collecting data from one household in Interview and Diary, respectively, and B_Y is the annual data collection budget. Using the method of Lagrange multipliers, we find that the optimal sample sizes are

$$n_I = \frac{\gamma B_Y}{c_D + c_I \gamma} \quad \text{and} \quad n_D = \frac{B_Y}{c_D + c_I \gamma}, \quad \text{where} \quad \gamma = \frac{\sigma_I}{\sigma_D}\sqrt{\frac{c_D}{c_I}}.$$

2.2 Estimating Data Collection Costs

To estimate the per-household cost of collecting data in Interview and Diary, we only consider activities whose costs would change appreciably with a change in sample allocation. Because some costs are reported collectively for Interview and Diary, we have to split them between the two surveys. Table 2 summarizes the assumed cost allocations.

Table 2: Cost Allocations

Activity %	Allocated to Interview	% Allocated to Diary	Justification
CEQ Interviewing	100%	0%	All Interview
CED Interviewing	0%	100%	All Diary
Postage and Shipping	0%	100%	Interview responses are submitted electronically, but diaries are physically shipped
Reinterview	75%	25%	The reinterview sample for Interview is three times as large as the reinterview sample for Diary
Time of Interview Address Listing	50%	50%	During a given 12 month period, the number of new addresses in sample is about the same for Interview and Diary
Address Research	50%	50%	

About 80% of completed interviews are non-bounding, so we estimate the number of Interview households by multiplying the number of completed interviews by 0.8. The bounding interview or the initial interview that is not used in estimation is collected to "bound" the recall of respondents and thereby prevent telescoping (Neter and Waksberg 1984). Then we estimate the per household cost by dividing the data collection cost by the number of households. In the end, our estimate of the ratio of the per household data collection cost for Interview to that of Diary is $c_I/c_D = 1.355$. The optimal allocation ratio γ depends on c_I and c_D only through this ratio.

2.3 Sample Selection Methods

Before we describe the selection methods, we first provide some necessary background to the CE sample design. As mentioned previously, Interview selects a *sys* sample from a list of addresses. We consider the specific addresses on the frame as the second-stage sampling units where the units are housing units. The frame is sorted by variables related to geography, family size, owner/renter status, property value, and monthly rent. Then the first unit of each cluster or hit string is chosen at equal intervals where the interval length is equal to the sampling interval. The clusters of 24 units work with the rotating-panel design of Interview. When a panel completes its 5th interview, it is replaced by another unit from the same cluster. This replacement of units from the same cluster reduces the variance of year-to-year change because the units are replaced by other units that are similar to them. We assume they are similar because they come from the same cluster and because units in the same cluster should be similar with respect to the sort variables.

2.3.1 Unstratified Samples

With "unstratified" samples we draw an *srswor* from the overall universe of units. Some hit strings can have zero units selected from them, while other hits can have multiple units selected. Simulating with an unstratified sample design was used because it is easy to implement and provide an upper-bound reference for the other variances.

Figure 1 is a representation of the unstratified simulation. Each dot represents a unit and the vertical lines group units from the same hit string. The dark unit represents a unit

selected into the sample of the simulation. Figure 1 shows that a *srswor* sample design can select no units, one unit, or multiple units from a given hit.

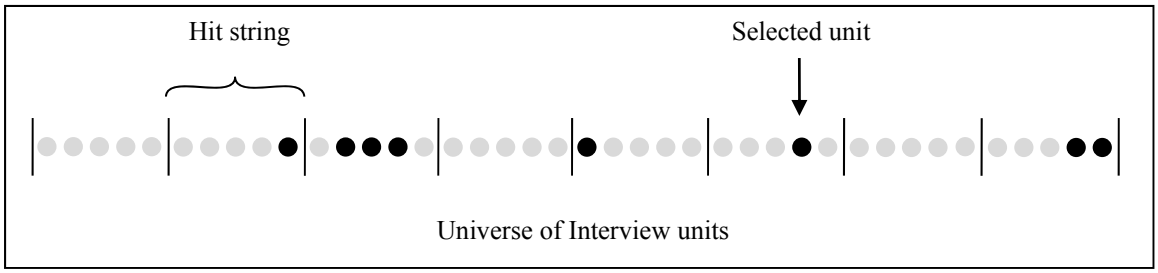

Figure 1: Example of an Unstratified Sample for Interview

2.3.2 Stratified Samples

To generate a "stratified" sample, we regard the original hit strings as strata and randomly select one unit from each hit string. If a selected unit has no associated households because of non-response, we add nothing to our sample. Using this process, we generate 50,000 stratified samples, each containing about 28,000 Interview households and 14,000 Diary households.

Figure 2 is a representation of the stratified simulation. The dots and lines of Figure 2 have the same meaning as in Figure 1. The stratified simulation is unique in that one unit is selected within each hit string.

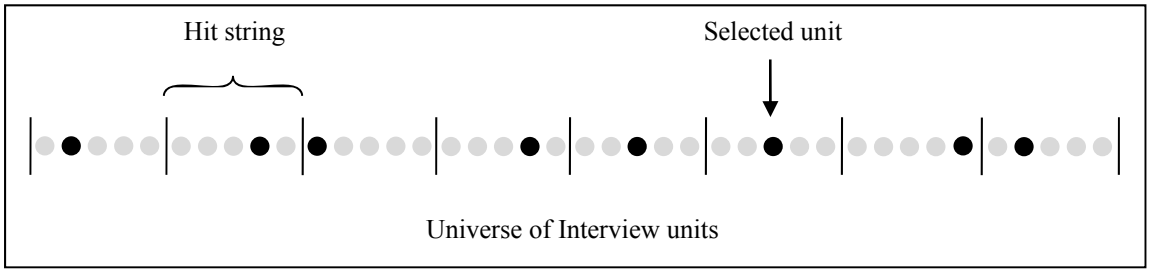

Figure 2: Example of a Stratified Sample for Interview

2.3.3 Yearly Samples

The last type of simulation uses the 4 samples from 2006 to 2009. With this method, the simulation is averaged over 4 samples instead of 50,000.

2.4 Estimating Variances

With each of the simulated samples from each of the methods, we estimated three different statistics: the mean expenditures $\overline{y}$, the *srswor* variance estimator, and the successive difference estimator. The simulation variance in Table 4 used ybar and calculated the simple variance of the mean expenditures over all simulations. The *srswor* and successive difference estimators averaged the value from each simulation over all simulations.

Table 3 displays our nine sets of variance estimates. The top and bottom numbers in each cell are estimates of the standard errors σ_I and σ_D, respectively.

Table 3: Standard Error Estimates

Variance Estimation Method	Sample Type		
	Stratified Un	stratified	Yearly
Simulation	$\hat{\sigma}_I$ = 26,656 $\hat{\sigma}_D$ = 10,815	30,053 11,284	18,709 2,096
SRSWOR	38,316 13,278	38,387 13,286	29,715 10,698
Successive Differences	38,800 13,147	38,640 12,163	27,233 10,525

Some observations about the standard errors in Table 3:

- The yearly-simulation estimates are unusually small. This has to do with our yearly inflation adjustments.
- As expected, the *srswor* method gives larger estimates than does simulation.
- The successive differences and *srswor* estimates are similar, but we thought the successive differences method would consistently produce smaller estimates. Using PSUs as strata may have increased our estimates because many successive differences are calculated between WHUs from different PSUs and frames.
- The optimal allocation ratio depends on the variance parameters through the ratio σ_I / σ_D. With the exception of the yearly-simulation estimates, all values of $\hat{\sigma}_I / \hat{\sigma}_D$ are close to 3.

2.5 Results

Table 4 summarizes the optimal allocation ratios. Disregarding the yearly-simulation ratio, all ratios fall in the range 2.12 – 2.73. Of the three variance estimation methods, simulation does the best job of taking CE's sample design into account. This method works best with the "stratified" and "unstratified" samples. The optimal ratios for these sets of estimates are 2.12 and 2.29, respectively.

Table 4: Optimal Allocation Ratios

Variance Estimation Method	Sample Type		
	Stratified Un	stratified	Yearly
Simulation 2.	12	2.29	7.67
SRSWOR 2.	48	2.48	2.39
Successive Differences	2.54	2.73	2.22

The last column in Table 5, SE Ratio, gives the ratio of the optimal standard error to the actual standard error to two decimal places. All values equal 1.00, which indicates the optimal allocation offers a very small improvement over the actual allocation.

Table 5: Allocation Results for the Stratified-Simulation Variance Estimates

Calendar Year	Optimal Allocation			Standard Error of $\overline{y}$		
	Interview	Diary	Ratio	Optimal	Actual	SE Ratio
2005 3	0,378	14,348	2.12	$560.86	$560.89	1.00
2006 2	9,316	13,847	2.12	$561.88	$561.90	1.00
2007 2	7,792	13,127	2.12	$563.47	$563.49	1.00
2008 2	8,184	13,312	2.12	$563.05	$563.09	1.00
2009 2	8,786	13,597	2.12	$562.41	$562.47	1.00

Figure 3 presents the values of the standard error of $\overline{y}$ as a function of the number of Interview households for the stratified-simulation variance estimates and a typical annual data collection budget. The minimum value occurs near n_I = 29,000 Interview households, but because the curve is so flat in the middle, our current annual sample size of 28,000 results in a standard error very close to the minimum. In fact, this is true for all values in the wide range 25,000 to 32,000.

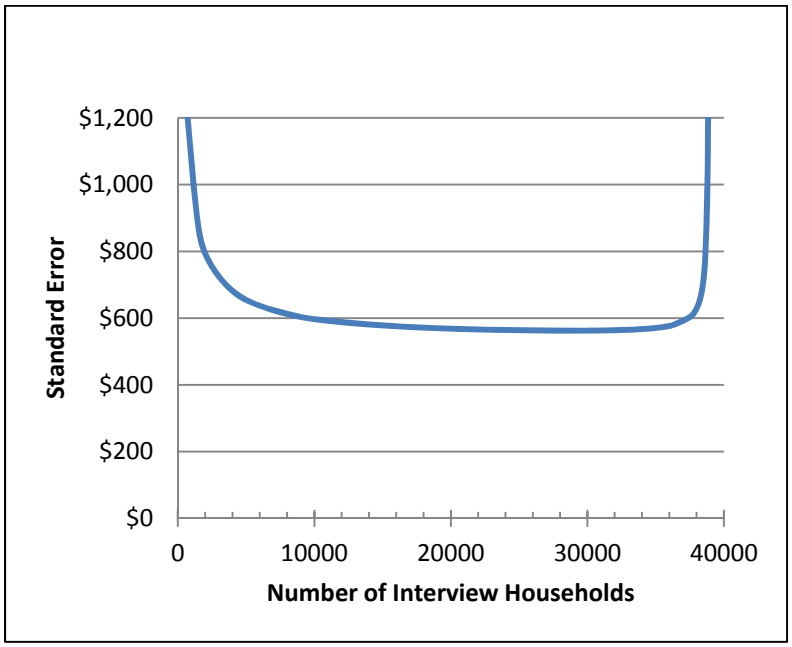

Figure 3: Standard Error of $\overline{y}$ as a Function of Number of Interview Households

Figure 4 presents the values of the standard error, this time expressed as a function of allocation ratio. You can see that the current ratio of 2.00 falls on the flat bottom of the curve.

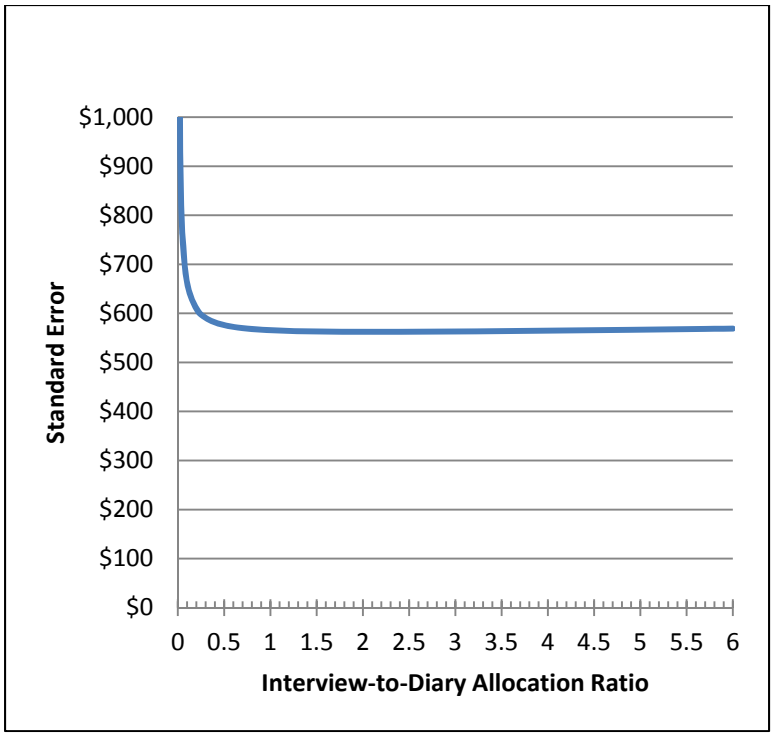

Figure 4: Standard Error of $\overline{y}$ as a Function of Allocation Ratio

Although our variance estimates produce different optimal ratios, they all result in a standard error curve that is very flat and thus very insensitive to changes in sample allocation. In every case, the standard error of $\overline{y}$ could be minimized by increasing the ratio, but the improvement over our current 2-to-1 allocation would be very small.

3. Allocation to the Second-Stage Sample Design

3.1 The research problem

The problem addressed by this research is what allocation of sample minimizes the second-stage design variance. We mention two important complications of the problem.

- There are many fine results available for allocating a sample to strata within a one-stage sample design; however, CE has a two-stage sample design.
- We are interested in the allocation for a given first-stage sample design since the first-stage sample design of CE only changes once every 10 years. We are not interested in allocating the sample to general stratum or a non-specific first-stage sample.

3.2 Notation

h	indexes the first-stage strata
i	indexes the first-stage units or PSUs of stratum *h*
j	indexes the counties of PSUs *i* in stratum *h*
k	indexes the units (households) of the second-stage sample design of given stratum *h*, and PSU *i*, and county *j*, respectively

$Y_h, \overline{Y}_h$	The total and the mean of *y* in stratum *h*
$Y_{hi}, \overline{Y}_{hi}$	The total and the mean of *y* in PSU *i* of stratum *h*
U	The overall universe of interest
s	The overall sample
U_{1h}	The first-stage universe of interest in stratum *h*
s_{1h}	The first-stage sample in stratum *h*
N_h	Number of units in stratum *h*
N_{hi}	Number of units in PSU *i* of stratum *h*
n_h	Number of second-stage sample units in stratum *h*
α_h	Stratum weight for stratum *h*, defined as $\alpha_h = N_h / N$
L	Number of first-stage strata

Estimators of statistics are denoted with a hat(^), for example, the estimators of the statistics $\overline{Y}_{hi}$ and $v_2\left(\hat{Y}_{hi}\right)$ are $\hat{\overline{Y}}_{hi}$ and $\hat{v}_2\left(\hat{Y}_{hi}\right)$, respectively. Overall, first-stage, and second-stage variances will be denoted as $v(.)$, $v_1(.)$, and $v_2(.)$, respectively.

3.3 Discussion of Variances of Interest

We are interested in finding an allocation method that best minimizes the variance of Interview. Let Y and $\overline{Y}$ be the overall total and the mean of the variable of interest *y*. Also let $\hat{Y}$ and $\hat{\overline{Y}}$ be the estimators of Y and $\overline{Y}$, respectively. Since CEQ has a two-stage sample design, the variance can be expressed as $v\left(\hat{\overline{Y}}\right) = v_1\left(\hat{\overline{Y}}\right) + v_2\left(\hat{\overline{Y}}\right)$ where the first-stage variance $v_1\left(\hat{\overline{Y}}\right)$ can be expressed generally as

$$v_1\left(\hat{\overline{Y}}\right) = \frac{1}{N}\sum_h \left[\sum_{i \in U_{1h}} \sum_{i' \in U_{1h}} \left(\pi_{hii'} - \pi_{hi}\pi_{hi'}\right) \frac{Y_{hi}}{\pi_{hi}} \frac{Y_{hi'}}{\pi_{hi'}} \right] \quad (1)$$

and the second-stage variance can be expressed as

$$v_2\left(\hat{\overline{Y}}\right) = \frac{1}{N^2}\sum_h \sum_{i \in U_{1h}} \frac{v_2\left(\hat{Y}_{hi}\right)}{\pi_{hi}}, \quad (2)$$

where *N* is defined as the size of the universe, π_{hi} is the probability of selection for PSU *i* in stratum *h*, π_{hij} is the probability of selecting PSU *i* and PSU *j* in stratum *h*. See Särndal *et al*. (1992; p. 137) for expressions (1) and (2).

Consider $v_1\left(\hat{\overline{Y}}\right)$ and note that it does not change unless the first-stage sample changes. Most importantly, it does not change no matter what n_h we choose for each first-stage

stratum. Since we have only observed the first-stage sample, we need to estimate $v_2\left(\hat{Y}_{hi}\right)$ with the sample we have as

$$\hat{v}_2\left(\hat{\bar{Y}}\right) = \frac{1}{N^2}\sum_h \sum_{i\in s_{1h}} \frac{\hat{v}_2\left(\hat{Y}_{hi}\right)}{\pi_{hi}^2} \tag{3}$$

where $m_h = 1$ for all strata h. The expression for (3) does change for different values of $\hat{v}_2\left(\hat{Y}_{hi}\right)$ will be bigger or smaller with decreasing or increasing sample sizes, respectively.

Table 6 summarizes the allocation methods considered by our research.

Table 6: Summary of Allocation Methods in Research

Method Lab	el	$n_{hi} \propto ...$
(1) Neyman-like Allocation	(1a)	$\alpha_h^2 v_2\left(\hat{\bar{Y}}_{hi}\right)$
	(1b)	$\left(N_{hi}S_{hi}\right)/\pi_{hi}$
(1c	)	$N_{hi}S_{hi}$
(2) Proportional to Size	(2a), (2b), (2c)	$N_h^{\frac{1}{2}}\ N_h\ N_h^2$
(2	d)	N_{hi}
(3) Constant Coefficient of Variation	(3a)	$\alpha_h\hat{\bar{Y}}_h$ or $N_h\hat{\bar{Y}}_h$ or $\hat{Y}_h$
	(3b)	$\hat{\bar{Y}}_h$
(3c	)	$\hat{Y}_{hi}$
(4) Equal Allocation	(4)	L (a constant)

The first set of allocation methods is similar to Neyman allocation in that they are function of the variance or a part of the expression for the variance. Method (1b) is a version of Neyman's allocation that we derived for the two-stage sample design. The derivation mimics Neyman's proof for one stage and the full proof can be found in Ash *et al*. (2012).

The second set of allocation methods is related to proportional to size of the strata or to the sample PSU within the strata. The third set assumes a constant coefficient of variation (the ratio of the standard error of the estimate and the estimate) for each strata or sample PSU and the last allocation method assumes an equal allocation for all strata.

Variances were estimated at two points of the research: before the allocation as input to the Neyman-like allocation methods and after the allocation in order to evaluate the alternative allocation methods. The variances were estimated from a set of 100,000 bootstrap samples selected from each county within each PSU of the current sample design. We used Interview completed interviews from 2004 to 2008, adjusted for inflation.

Table 7 summarizes the results of the simulation. The allocation methods are ordered by their coefficient of variation.

Table 7: Summary of Results

Relative Ratio					
All Methods	ok Methods	Coefficient of Variation	Method	Sample sizes for small PSUs are …	Proportional to…
1.000	1.000	0.53000	1b	ok	$\left(N_{hi}S_{hi}\right)/\pi_{hi}$
1.006	1.006	0.53344	3a	ok	$\hat{Y}_h$
1.009		0.53498	1a	too small	$\left(N_{hi}S_{hi}\right)^2$
1.024	1.024	0.54261	2b	ok	N_h
1.104	1.104	0.58487	2a	ok	$N_h^{\frac{1}{2}}$
1.152	1.152	0.61066	2c	ok	N_h^2
1.212	1.212	0.64259	3b	ok	$\hat{\bar{Y}}_h$
1.297	1	0.68766	4	ok	L
2.025		1.07341	2d	too small	N_{hi}
2.324		1.23168	1c	too small	$N_{hi}S_{hi}$
2.400		1.27191	3c	too small	$\hat{Y}_{hi}$

The fourth and sixth columns of Table 7 identify the allocation methods. The third column provides the coefficient of variation from the simulation and the first and second columns, the relative ratios, compare the coefficient of variation of the different methods. The relative ratio is the ratio of the coefficient of variation for the given method with the coefficient of variation of the method with the smallest coefficient of variation of the given set of methods.

Two sets of minimum points. A surprising finding of the research is that there appears to be two sets of allocation methods that minimize the variance. The fifth column of Table 7 identifies the two sets. There is a set of allocation methods, which did well, but allocated most of the sample to the large PSUs and only the minimum to the small PSUs. We refer to this set of allocation methods as "too small" in Table 2 since their sample sizes for the small PSUs were too small. We consider the "too" small methods as unacceptable. Our initial criterion of minimum overall variance is too simple and does not address the need for the Interview survey to have adequate representation in all PSUs. We therefore selected the recommended allocation methods from the set of allocation methods that had an adequate sample size for all PSUs, which we refer to as the "ok" methods in Table 2. The second column of Table 2 compares the "ok" methods with each other.

The smallest variance is associated with methods (1b) and (3a). This is not unexpected as the two methods use information about the variance of the strata. The next best allocation methods were (2b), (2a), and (2c), which used information about the size of the strata. The worst method was method (4) or constant allocation, which assumed nothing about the strata.

Not unexpectedly, the two best methods, (1b) and (3a), are both proportional to the variance of an estimated total. For (1b), the total is $\hat{Y}_{hi}$ and for (3a) the total is $\hat{Y}_h$.

4. Summary

We recommend using the current 2-to-1 Interview-to-Diary allocation ratio and allocating sample to PSUs proportional to $\left(N_{hi}S_{hi}\right)/\pi_{hi}$ or $N_h\hat{\bar{Y}}_h$.

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.

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the U.S. Bureau of Labor Statistics.

References

Ash, S., Croos, J., Swanson, D., Reyes-Morales, S., (2012). "Sample Allocation Research for the Consumer Expenditure Interview Survey," dated March 16, 2012.

Dumbacher, B., Swanson, D., Reyes-Morales, S., Steinberg, B., Ash, S. (2011). U.S. Census Bureau and Bureau of Labor Statistics Research Paper "Optimizing the Sample Allocation for the Consumer Expenditure Surveys," dated July 26, 2011.

Cochran, W.G. (1977). *Sampling Techniques*, John Wiley & Sons.

Neter, J. and Waksberg, J. (1964). "A study of response errors in expenditures data from household interviews," Journal of the American Statistical Association, 59, 18-55.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection, *Journal of the Royal Statistical Society*, 97, 558-625.

Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag.

Swanson, D. (1999). Bureau of Labor Statistics Research Paper "Optimizing the CE's Diary/Interview Sample Allocation," dated September 16, 1999.