Exploring Statistical Issues of Annual Sampling for the Current Population Survey

Reid Rottach, Antoinette Lubich, Benjamin M. Reist
US Census Bureau, Demographic Statistical Methods Division, 4600 Silver Hill Road,
Washington DC 20233

Abstract
The Current Population Survey (CPS) sample selection has been based on Decennial Census data, so sampling has been done once every ten years. With the introduction of the American Community Survey (ACS) and a semiannually updated Master Address File (MAF), and with greater access to administrative records, it has been proposed that CPS sample selection be done more frequently. In this paper we discuss a general method of selecting second-stage CPS sample annually, which maintains the longstanding 4-8-4 sampling scheme. We address the expected benefits as well as drawbacks of changing to annual sampling. Specifically, this will include a discussion of the expected effect on changes in between-month correlations on CPS estimates and the efficiency of CPS estimates.

Key Words: dynamic sampling, second-stage sample selection, 2010 Sample Redesign

1. Introduction

The 2010 based Sample Redesign of the Current Population Survey (CPS) is being planned to take advantage of the information available from the 2010 Decennial Census, the American Community Survey (ACS), administrative records, and the continuously updated Master Address File (MAF). A sponsor of the CPS has expressed an interest in annual sampling for the 2010 Redesign. The opportunity to redesign survey samples throughout the decade now exists. There would still be a sample redesign, possibly every five years, to perform the usual first-stage sample selection of geographical Primary Sampling Units (PSUs), to accommodate redefined survey objectives, to introduce new methodologies, etc. However, the second-stage selection of CPS sample design (selection of housing units (HUs)) would be an annual event. (For a detailed overview of the CPS design, refer to Chapter 3, 'Design of the Current Population Survey Sample,' of *CPS Technical Paper 66* listed in the Supporting Material section of this paper.)

The goal of this paper is to provide the plan of research for measuring the impact on data quality of selecting annual samples of HUs from the MAF, within the sampled PSUs. Also provided are preliminary results, a summary, and continuing CPS research.
.

1.1 Benefits of Annual Sampling

With the lapse in time since the last census, the efficiency of survey samples decline. For example, many sample addresses are lost or difficult to locate

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

because of demolitions and conversions to nonresidential use. Sending interviewers out into the field to look for addresses that no longer exist or are difficult to locate because of out-of-date HU lists may increase the overall data collection costs, sampling errors and nonsampling errors.

In addition to lessening the difficulties mentioned in the paragraph above, the following are other advantages for periodic/annual sampling of HUs from extracts of the MAF. The relevance of some of them can be given in terms of increased efficiency. Basically, a more efficient sample is one which either lowers variance for a given sample size, or results in a smaller sample size for the same variance. Note that sample size is positively correlated to cost; i.e., cost increases as the sample size increases.

- Switching to more frequent sampling would eliminate the continuous cycle of increasing sample sizes (i.e., increasing costs) with growth followed by maintenance reductions. Sampling annually would capture any needed increases.
- Because the method of annual sampling would no longer be reliant on reduction groups or reserve/research samples, it could be easier to implement a change in sample to meet changes in budget, priorities, and population over the design.
- Annual sampling presents the option of using Decennial or ACS data, with administrative records, choosing whichever provides the most up-to-date data for the second-stage sample selection. ACS estimates will be available as 5-year rolling averages, so the benefits of using ACS data will be the greatest in the latter years of the design, as one gets further from the release of the Decennial Census data. The option to choose the most current data would produce a more efficient sample.
- Annual sampling would mean no more skeleton frame[i] sampling for new construction. (Note: A roman numeral superscript indicates an endnote.) Skeleton frame sampling results in less predictable sample sizes and less efficient stratifications.

1.2 Drawbacks of Annual Sampling

The following are several cons of annual sampling. The first two are in reference to Figure 1, 'CPS 4-8-4 Sampling Every Year' in Section 6.1.

- There will be either two or three sample selections in any given month. Any one sample selection will be at most 50 percent of the entire monthly sample. Thus, the national and state sampling intervals would change monthly to accommodate a fixed sample size, which could result in more variation in estimation weights.
- The problem of respondent burden will have to be addressed. Currently a HU can only be selected once in a design.
- Flexibility in adding or removing sample could affect FR hiring and planning.

2. Background

2.1 MAF and DSF Defined

The MAF is intended to be a nationwide list of all living quarters with their geographic locations. The original file was created by combining the addresses in the 1990 address control file with the U.S. Postal Service (USPS) delivery sequence file (DSF), and supplementing this with address information provided by state, local, and tribal governments, and Census field operations. The MAF is the sampling frame for the ACS.

The DSF is a nationwide computer address file of all residential and commercial units that receive mail delivered by the USPS. It provides information about existing addresses, new addresses, changes in the status of existing addresses, and demolished HUs, and is the most current national source of addresses used for mail delivery. The Census Bureau uses the DSF as a source for maintaining and updating its MAF[ii]. Although the USPS usually provides customers with a monthly file of DSF updates, the Bureau has neither the resources nor the need for monthly updates. It currently receives a DSF in March and September to account for all DSF changes occurring over the previous six months.

2.2 CPS 2000 Redesign Sample Selection Aspects Applicable to this Research

The CPS sample is a multi-stage stratified sample of approximately 72,000 assigned HUs from 824 sample areas, designed to measure demographic and labor force characteristics of the civilian noninstitutionalized population 16 years of age and older (CNP16+). The CPS samples HUs from lists of addresses obtained from the 2000 Decennial Census. The sample is updated continuously for new housing built after Census 2000.

The first stage of sampling involves dividing the U.S. into PSUs, most of which comprise a metropolitan area, a large county, or a group of smaller counties. After grouping the PSUs into strata, one PSU is sampled in each stratum.

The second stage of the CPS sample design is the selection of sample HUs within the sample PSUs. As stated in *CPS Technical Paper 66*, to accomplish the objectives of within-PSU sampling, data from the 2000 Decennial Census and the Building Permit Survey are used. These two sources, and the relationship between them, are used to develop the four sampling frames: the unit frame, the area frame, the group quarters (GQ) frame, and the permit frame. The unit, area, and GQ frames are collectively called old construction, and the permit frame represents new construction[iii].

In the second stage of sampling, for the unit frame, HUs within a Basic PSU Component (BPC)[iv] are sorted based on geographic and demographic characteristics[v]. After sorting, a systematic sample is drawn to create clusters of 84 HUs. These clusters are called hit strings. Each contains 21 hits of 4 HUs. Since there is one hit for each of the 21 sample designations, this ensures that one hit from each hit string is in sample in every month. To reduce disclosure risk, we also ensure adjacent HU's are not in sample at the same time. (Figure 1 in Section 6.1 illustrates 3 sample designations.)

Sample is partitioned into eight representative subsamples called rotation groups. The eight subsamples are balanced across stratification PSUs, states, and the nation. Rotation group is used in conjunction with sample designation to determine units in sample for particular months during the decade. Units are in sample for 4 months, rest for 8 months, and return to sample for 4 months. This is the basis of the CPS 4-8-4 rotation scheme.

Overall, the objectives of the 2000 Sample Redesign within-PSU sampling were to:

- Select a series of probability samples that are representative of the CNP16+.
- Give each HU in the population one and only one chance of selection, with all HUs in a state or substate area having the same overall chance of selection.
- Keep the within-PSU variance on labor force statistics (in particular, unemployment) at as low a level as possible, subject to respondent burden, cost, and other constraints.
- Put particular emphasis on providing reliable estimates of monthly levels and change over time (e.g., year-over-year) of labor force items.

3. Assumptions and Out-of-Scope

The following are basic assumptions for the CPS research described in this paper:

- Except for the method and timing of selecting HUs in the second stage, the basics of the CPS survey design will still be in effect; e.g., a 4-8-4 rotation design with rotation groups.

- There may or may not be some form of an area/'hybrid' area frame; however, although this research and its findings are based on single-frame MAF, they are believed to be valid regardless of the form of the area frame.
- There will be MAF updates with a DSF twice a year.
- The ACS and administrative records will be used for demographic sort/stratification variables in the second-stage sample selection of HUs[vi].
- Because of required work, the approval process, and interactions with other branches/divisions/sponsor, sample selection must be completed six months prior to the date for planned implementation.
- Sampling will be done independently for each PSU.
- Estimates will be unbiased.

The following are out-of-scope for the CPS research described in this paper:

- Demographic surveys other than the CPS.
- The use of the ACS as a possible sampling frame.
- Accommodations for new surveys and special surveys.
- Oversampling.
- Operational issues.
- Confidentiality concerns.
- Unduplication of HUs across demographic surveys and with the ACS.
- Costs of implementation and maintenance, including possible increased listing costs depending on the existence of some form of an area frame. These costs will be a significant deciding factor and should prove to be a major focus of 2010 redesign efforts. However, costs are indirectly dealt with in the results and summary sections since, based on variance calculations, switching to annual sampling would require a change in the current sample size.
- Transition from the sampling system used in the 2000 design.

4. Survey Requirements

Chapter 3, 'Design of the Current Population Survey Sample' of *CPS Technical Paper 66*, lists five survey requirements and they are summarized below. If we implement annual sampling, we will continue to adhere to these requirements.

- The CPS sample is a probability sample.
- The sample is designed primarily to produce national and state estimates of labor force characteristics of the CNP16+.
- The CPS sample consists of independent designs for the states and substate areas; it is state-based.
- The specified CV reliability requirement for the monthly unemployment (UE) level for the nation, given a 6 percent UE rate, is 1.9 percent or less. This follows from the requirement that a difference of 0.2 percentage points in the UE rate for two consecutive months be statistically significant at the 0.10 level.
- The required CV on the annual average UE level for each state, substate, and the District of Columbia, given a 6 percent UE rate, is 8 percent.

Related to the fourth bullet, if the correlation structure changes via the annual sampling method, we will have to determine the CV reliability for the UE level for the nation that we should control to, given a 6 percent UE rate and given the requirement that a difference of 0.2 percentage points in the UE rate for two consecutive months be statistically significant at the 0.10 level.

5. Research Questions

5.1 What is the Effect of the Change in Between-Month Correlation on CPS Estimates Caused by Annual Sampling?

The largest component of between-month correlation in CPS estimates is caused by overlapping sample between months induced by the 4-8-4 rotation scheme. Since annual sampling will maintain this rotation scheme, this source of correlation between months will be unchanged under annual sampling. Use of hit strings in the current design also results in an additional, though smaller, component of between-month correlation in CPS monthly estimates. Hit strings ensure that the hits that rotate out of sample are replaced by hits from the same hit string. Since HUs in a hit string have geographic and demographic characteristics in common, it is thought that hits in a hit string are correlated. Annual sampling will eliminate hit strings and thus this smaller component of between-month correlation.

A reduction in correlation would affect two major types of CPS estimates: estimates of variance on month-to-month change and year-over-year change estimates; and estimates of variances on annual average estimates.

5.1.1 Estimates of variance on month-to-month change and year-over-year change estimates

In general, lower positive correlation between months will increase the variance on changes between months. In the CPS, we are particularly interested in the month-to-month and year-over-year changes in national estimates of national civilian labor force (CLF), UE, and employment (EMP) levels. We are also interested in the month-to-month and year-over-year changes in the national UE rate. The detectable difference between these from month-to-month and year-over-year could be significant, depending on the size of this change in correlation between months.

5.1.2 Estimates of variance on annual average estimates

In general, lower positive correlation between months will decrease the variances on annual average estimates. In the CPS we are particularly interested in the state annual average estimates of CLF, UE, and EMP levels. We are also interested in the state annual average unemployment rate. Depending on the size of the change in correlation, we could see a significant decrease in estimates of variance on annual average estimates.

5.2 What is the Effect of Annual Sampling on the Efficiency of CPS Monthly Estimates?

There are two areas in which annual sampling might improve the efficiency of CPS monthly estimates. The first is that since the ACS and administrative records will provide new demographic estimates annually, each annual sample will have the benefit of being sorted on more current demographic sort/stratification variables. In the 2000 design, sample was selected once, based on census information. The second is that new construction will be sorted along with old construction. In the 2000 design, new construction was sorted with a less detailed sort than the unit frame.

5.3 How will Unduplication across Sample Selections and the Inclusion of MAF Adds[vii] Between Sample Selections be Handled?

In the 2000 design, unduplication within CPS sample did not have to be addressed since a HU could only be selected once for the life of the design. With annual sampling, a method will need to be devised to a keep HU from being sampled too frequently (e.g., more than once every five years). Depending on the unduplication method that is chosen, there could be implications for variances of key CPS estimates and/or calculating conditional probabilities of selection. In terms of the inclusion of MAF adds that occur between sample selections, they will be sorted and selected along with old construction HUs under annual sampling. This will make calculating conditional probabilities of selection more complicated due to the without replacement[viii] strategy.

6. Methods

In this section, since annual sampling will not affect the first-stage selection of PSUs, this component is ignored. Variances, for example, are assumed to be those due to sampling within PSUs.

6.1 Simulating the Annual Sampling Correlation Structure and Its Effect on CPS Estimates

We expect the correlation structure of our sample to change with annual sampling, which can have a positive effect on some variances and a negative effect on others. Two situations are described:

- With our current approach, new HUs that enter sample are likely to be similar to HUs already in sample due to the method of sampling HU clusters (hit strings), defined on geographic and demographic variables. We would likely lose some of this uniformity of our sampled HUs with annual sampling. Generally speaking, clustering increases variances, so we might observe lower variances for monthly estimates.
- A possible drawback to annual sampling would be lower between-month correlations since HUs in one annual sample are selected independently[ix] of HUs in another annual sample, rather than belonging to common clusters. Lower between-month correlations in our estimates would in turn lead to higher variances in estimates of change.

We would like to simulate the effect that annual sampling will have on correlations, using data from the 2000 design. Our variance estimator inherently accounts for the geographic clustering by assigning all HUs in a hit string the same variance cluster code. If we assigned HUs within the same hit string to different variance clusters, we can remove the effect of clustering from our variance estimates. Although the variance estimator is biased given our true sample design, it gives an idea of variances we can expect to achieve in the alternate design. To simulate annual sampling, we will redistribute HUs among variance clusters in a way that is reflective of our proposed annual sampling plan. Figure 1 illustrates the approach to annual sampling that will guide the simulation.

6.1.1 Phase in of annual samples

From the rotation chart on the next page, monthly estimates use data from eight rotation groups represented in a single row. The chart is split into two Sample Selection (SS) groups as an illustration; these are two independently selected annual samples. Monthly estimates use data from either two or three SS groups. Any monthly estimate after the phase in of the first SS will contain four rotations from one of the SS groups, with the remaining four rotations distributed in one or two remaining groups. Furthermore, the overlap in sample between two months, represented by the overlap between two rows on the chart, will be the same as in the old design. For example, between consecutive months, there will be six of eight rotation groups in common; i.e., a 75 percent HU overlap. What will change is the correlation structure between rotation groups in different annual samples.

		Sample Designation 1 (SD 1)								*Sample Designation 2 (SD 2)*								*Sample Designation 3 (SD 3)*							
		1st Sample Selection (1st SS)												*2nd Sample Selection (2nd SS)*											
Year 1	*Jan*	**1**																							
	Feb	**1**	**2**															**Phase-In 1st SS**							
	Mar	**1**	**2**	**3**																					
	Apr	**1**	**2**	**3**	**4**																				
	May		**2**	**3**	**4**	**5**																			
	Jun			**3**	**4**	**5**	**6**																		
	Jul				**4**	**5**	**6**	**7**																	
	Aug					**5**	**6**	**7**	**8**									**50% 1st SS In Sample**							
	Sep						**6**	**7**	**8**	**1**															
	Oct							**7**	**8**	**1**	**2**														
	Nov								**8**	**1**	**2**	**3**													
	Dec									**1**	**2**	**3**	**4**												
Year 2	*Jan*	**1**									**2**	**3**	**4**	**5**											
	Feb	**1**	**2**									**3**	**4**	**5**	**6**			**Phase-In 2nd SS**							
	Mar	**1**	**2**	**3**									**4**	**5**	**6**	**7**									
	Apr	**1**	**2**	**3**	**4**									**5**	**6**	**7**	**8**								
	May		**2**	**3**	**4**	**5**				**50% 1st SS**					**6**	**7**	**8**	**1**							
	Jun			**3**	**4**	**5**	**6**			**and**						**7**	**8**	**1**	**2**						
	Jul				**4**	**5**	**6**	**7**		**50% 2nd SS**							**8**	**1**	**2**	**3**					
	Aug					**5**	**6**	**7**	**8**									**1**	**2**	**3**	**4**				
	Sep						**6**	**7**	**8**	**1**									**2**	**3**	**4**	**5**			
	Oct							**7**	**8**	**1**	**2**									**3**	**4**	**5**	**6**		
	Nov								**8**	**1**	**2**	**3**									**4**	**5**	**6**	**7**	
	Dec									**1**	**2**	**3**	**4**									**5**	**6**	**7**	**8**
Year 3	*Jan*										**2**	**3**	**4**	**5**									**6**	**7**	**8**
	Feb	**Phase-Out 1st SS**										**3**	**4**	**5**	**6**									**7**	**8**
	Mar												**4**	**5**	**6**	**7**									**8**
	Apr													**5**	**6**	**7**	**8**								
	May														**6**	**7**	**8**	**1**							
	Jun															**7**	**8**	**1**	**2**						
	Jul																**8**	**1**	**2**	**3**					
	Aug	**50% 2nd SS In Sample**																**1**	**2**	**3**	**4**				
	Sep																		**2**	**3**	**4**	**5**			
	Oct																			**3**	**4**	**5**	**6**		
	Nov																				**4**	**5**	**6**	**7**	
	Dec																					**5**	**6**	**7**	**8**
Year 4	*Jan*																						**6**	**7**	**8**
	Feb	**Phase-Out 2nd SS**																						**7**	**8**
	Mar																								**8**

Figure 1: CPS 4-8-4 Sampling Every Year (Two sample selections shown.)

6.1.2 The simulation

We assume the correlation structure among rotation groups within an annual sample will be similar to what we observed in the 2000 design, but the correlation structure between rotations in different annual samples will change. To simulate this, the new rotation groups entering sample for the first twelve months will not have their variance codes changed, but the following twelve rotation groups will have their codes changed so that HUs in a given hit will remain in the same variance cluster, but hits within the same hit string will be randomly redistributed into different clusters. Similarly, we can do this the following year to simulate a third annual sample.

In practice, the variance codes are replicate factor assignments which we force to be uncorrelated between HUs moved to different clusters. (For further details, refer to Fay (1984, 1989).) The replicate factors can be run through all the steps of CPS replicate weighting. With these newly created replicate weight files, we can compute annual sampling variances and correlation matrices for CPS statistics estimated with second-stage or composited weights. These variances and correlations can be compared to those of the true sample design.

Estimating Within-PSU Correlations for the Annual Sampling Simulation:

Consider two monthly Horwitz-Thompson estimators, which are the sum of hit level totals:

$\hat{X} = \sum_{k=1,\ldots,n} x_k$ is an estimated total for the first month, and

$\hat{Y} = \sum_{k=1,\ldots,n} y_k$ is an estimated total for the second month.

For the current design, the *k*th element in the first month is in the same hit string as the *k*th element in the second month.

The successive difference replication estimator of covariance can be written as:

$$Cov_{SDR}\left(\hat{X},\hat{Y}\right) = \left[x_1 \;\; x_2 \;\; \ldots \;\; x_n\right]\mathbf{C}\begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix}$$

where **C** is one-half times a matrix with "2"s on the diagonal, "-1"s in the positions adjacent to the diagonal, and zeros everywhere else. This leads to an explicit form for the covariance of: $Cov_{SDR}\left(\hat{X},\hat{Y}\right) = \frac{1}{2}\left(x_1 y_1 + \sum_{k=1}^{n-1}\left(x_k - x_{k+1}\right)\left(y_k - y_{k+1}\right) + x_n y_n \right)$.

In the case of annual sampling, many terms in the sum for covariance would have expected value zero. To simulate this, we identified those terms each month and effectively modified the **C** matrix so that the estimator of covariance had those terms removed. That is, if element *k* is in a different annual sample, the *k*th row and column of **C** is replaced with zeros.

In the case where an entire rotation group is in a new annual sample, the matrix **C** becomes block diagonal with the block from the new rotation group being a zero-matrix, and everything else having the same pattern of "2"s and "-1"s as before. So the explicit form of the covariance would have removed all terms involving units in that rotation.

The expressions above are equivalent to the form involving replicate factors (see Fay and Train (1995) and Fay (1984, 1989). In our application, the zeros are obtained by assigning orthogonal rows of a Hadamard matrix when constructing the replicate factors. That is, if unit *k* in the second month is assigned Hadamard row vectors that are orthogonal to the row vectors assigned to unit *k* in the first month, Fay's replication estimator becomes equivalent to that in which **C** has zeros in the *k*th row and column. Fay's replication estimator has the form $Cov_{SDR}\left(\hat{X},\hat{Y}\right) = \frac{4}{160}\sum_{r=1,\ldots,160}\left(\hat{X}_r - \hat{X}\right)\left(\hat{Y}_r - \hat{Y}\right)$

where $\hat{X}_r$ and $\hat{Y}_r$ are the estimates using replicate weights. We used a perturbation factor of 0.5.

Although the previous results describe a covariance estimator for Horwitz-Thompson statistics, we followed Fay's approach to compute covariances on reweighted estimates. Furthermore, for each lag, we computed a covariance for each pair of months separated by that lag, and then computed a correlation. The correlations presented in this paper are averages across all the monthly pairs.

6.2 Measuring the Gain in Efficiency of CPS Estimates Due to Improved Stratification

With annual sampling, every year we will select a new second-stage sample, stratifying on the most recent data rather than only data from year 2010. This may lead to gains in efficiency in monthly estimates in subsequent years since the stratification variables will be updated. Furthermore, we will be able to improve the stratification of new construction. In the past, we have had very little information to stratify on when this sample was selected. One possible way to quantify the loss in efficiency with time would be to examine the behavior of our design effects as the stratification variables aged, and as new construction made up a higher portion of the 2000 sample. If the design effects increase noticeably, we would assume it was largely due to inefficient stratification that can be overcome by more frequent sampling.

The X-12-ARIMA[x] and PROC ARIMA[xi] tools are used to answer this research question. Monthly CLF, UE, and EMP level data from July 2005 (when phase-in of the 2000 redesign was complete) through March 2009 were available for analysis. Design effects for each of these three variables, for each of the 45 months, after both the second-stage of weighting and after compositing, were calculated and used as the input.

6.3 Theoretical Issues of MAF Adds and Unduplication Across Sample Selections

With annual sampling, we would like to maintain predetermined probabilities of selection, and address the issue of dependence between sample selections that may arise when we ensure HUs are selected without replacement. We will discuss different approaches, their effects on probabilities of selection, variances and covariances.

7. Preliminary Results

7.1 Simulating the Annual Sampling Correlation Structure and Its Effect on CPS Estimates

7.1.1 The approximate variance of a correlation

Fisher's Z transformation allows us to approximate the variance of a correlation from the assumption that the covariance matrix follows a Wishart distribution (*W*). This belongs to multivariate normal theory, in which case *W* is the distribution of the *"n-1"* sample covariance estimator. It is sometimes assumed the covariance estimator we use follows the same distribution, even though our approach to sampling is not SRS, and the form of the covariance estimator is a different function of the data. Our replication covariance estimator for an estimated vector of totals $1'\mathbf{X}$ can be expressed as $\mathbf{X}'\mathbf{C}\mathbf{X}$. We assume in this case that $\mathbf{X}'\mathbf{C}\mathbf{X} \sim W(k, \Sigma)$, where $\mathbf{X}'\mathbf{C}\mathbf{X}$ is the estimated covariance matrix, k is a degrees of freedom parameter, and Σ is related to the expected value of $\mathbf{X}'\mathbf{C}\mathbf{X}$ as $E[\mathbf{X}'\mathbf{C}\mathbf{X}] = k\Sigma$.

Fisher's Z transformation depends on the covariance matrix only through the sample correlation, as $Z = \frac{1}{2}\log\left(\frac{1+\hat{\rho}}{1-\hat{\rho}}\right)$. Fisher showed this statistic has an approximate Normal distribution with variance $\frac{1}{k-3}$. The inverse of Z is $\hat{\rho} = \frac{e^{2Z}-1}{e^{2Z}+1}$, which leads to a linearized variance estimator for the sample correlation given by $\hat{V}(\hat{\rho}) = \frac{\left(1-\hat{\rho}^2\right)^2}{k-3}$.

For a linear statistic, the rank of **C** in our estimator **X′CX** equals the number of rows of a Hadamard matrix we used to deconstruct **C**. (For a description of our variance estimator, see Fay (1984, 1989).) The degrees of freedom will be less than or equal to the rank of **C**, and we assume that the upper bound is reached for statistics representing large areas, such as large states or the nation. In our case, the upper bound is 158 degrees of freedom.

So, the standard error of a sample correlation based on our replication variances is approximately $s\hat{e}(\hat{\rho}) = .08\left(1 - \hat{\rho}^{2}\right)$.

7.1.2 Comparing average correlations

We estimate correlations using 25 months of data, having 24 estimates of lag 1 correlation, 23 estimates at lag 2, 22 estimates at lag 3, and so on. For each lag, the estimates will be treated as being independent.

In significance testing for the difference in correlations between the current sampling approach and the simulation, we assumed there were correlations-between-the-correlations related to the proportion of rotations with identical variance code assignments. For example, in August 2006, seven of eight rotations had identical assignments for the current method and the simulation. Across all 25 months, there were 68 rotations with identical assignments out of 200, so we approximate the correlations at 34 percent.

The standard error of the differences is assumed to be $s\hat{e}(\overline{\rho}_c - \overline{\rho}_s) = \dfrac{\sqrt{\operatorname{var}(\hat{\rho}_c) + \operatorname{var}(\hat{\rho}_s) - .68\, se(\hat{\rho}_c) se(\hat{\rho}_s)}}{\sqrt{25 - nlag}}$, where the subscripts *c* and *s* are the current approach and the simulation, *nlag* is the lag size, and the average correlations are used in the variance formula, rather than any individual correlation estimates.

7.1.3 Sample size

In Table 1, there appears to be a small decrease in correlation due to annual sampling. The CPS selects enough sample to allow a 0.2 percent difference in the UE rate to be significant at the 10 percent level. So, a drop in month-to-month correlation would require a larger sample size. If we only consider the effect of changing within-PSU correlations, the sample size requirement can also be expressed as requiring the within-PSU variance of a difference remain constant in the new design. Assuming the within-PSU variance is inversely proportional to the number of HUs selected, this leads to

$$\frac{n_s}{n_c} = \frac{1-\rho_{w,s}}{1-\rho_{w,c}}.$$

Using the estimated correlations, this ratio is 1.05. Thus, we would expect to need a 5 percent increase in sample size if we were to switch to annual sampling, unless we could distribute the additional sample more effectively across the states.

Table 1: Within-PSU Correlations

	Unemployment[xii]		*Employment*		*Civilian Labor Force*	
Lag	*Annual sampling*	*Current approach*	*Annual sampling*	*Current approach*	*Annual sampling*	*Current approach*
1	0.330*	0.361	0.595*	0.630	0.590*	0.616
2	0.168	0.183	0.362*	0.406	0.363	0.394
3	0.073	0.085	0.180*	0.224	0.190	0.219
4	0.012	0.011	0.032	0.062	0.035	0.046
5	0.024	0.028	0.044	0.057	0.048	0.037
6	-0.029*	0.033	0.051	0.057	0.053	0.043
7	-0.004	0.039	0.049	0.047	0.048	0.034
8	-0.024*	0.040	0.078	0.060	0.071	0.037
9	0.004*	0.056	0.150	0.128	0.150	0.111
10	0.011*	0.064	0.193	0.195	0.195	0.164
11	-0.016	0.034	0.240	0.249	0.248	0.228
12	0.003*	0.077	0.235*	0.292	0.256	0.283
13	0.032	0.076	0.157*	0.233	0.171	0.212
14	-0.004*	0.058	0.106*	0.170	0.110	0.143
15	-0.033*	0.064	0.067	0.106	0.070	0.090
16	-0.035*	0.054	0.047	0.071	0.048	0.053

* Significantly different from the current approach at the 10 percent confidence level.

Solving for the needed change in sample size for annual sampling:

$V\left(\hat{X}-\hat{Y}\right)\cong 2V(\hat{X})(1-\rho)$

If we assume $V(\hat{X})=\frac{\sigma^2}{n}$, and compare the variance of a difference for the current approach to that of annual sampling, we assume the σ^2 will be the same, but the correlations will be different. In order for the variance of a monthly change to be the same under annual sampling as it is for the current design, we will have to adjust sample sizes.

The relationship $2\frac{\sigma^2}{n_c}(1-\rho_c)=2\frac{\sigma^2}{n_a}(1-\rho_a)$ leads to $\frac{n_a}{n_c}=\frac{(1-\rho_a)}{(1-\rho_c)}$.

The assumption that variance is inversely proportional to sample size:

This relationship holds for simple random samples, but we cannot immediately assume it would be true in our case. In the last design when the number of eligible HUs in sample was reduced from 56,000 to 50,000, the CV on UE changed from 1.8 percent to 1.9 percent. Therefore, $\frac{V_1(\hat{X})}{X^2}\Big/\frac{V_2(\hat{X})}{X^2}=\left(\frac{1.9}{1.8}\right)^2$. The left side of this equation reduces to the ratio of variances and the right side is approximately the ratio of sample sizes (56,000/50,000), which is consistent with our assumption.

7.2 Measuring the Gain in Efficiency of CPS Estimates Due to Improved Stratification

The CLF, EMP, and UE design effects, through the second stage of weighting, from August 2005 through March 2009 were plotted. It was decided to check for seasonality and then to try to fit a time series model for each.

When the response series has a seasonal pattern, the values of the series at the same time of year in previous years may be important for modeling the series. All three sets of design effects exhibited no seasonal patterns. We saw no additional need to test for seasonality on composited data.

The white noise check is an approximate statistical test of the hypothesis that none of the autocorrelations of the series up to a given lag are significantly different from zero. If this is true for all lags, then there is no information in the series to model, and no ARIMA model is needed for the series. This is exactly the result we obtained using the design effects through the second stage of weighting. The p value for the test of the first six autocorrelations (Pr>ChiSQ) was 0.7257 for CLF, 0.8376 for EMP, and 0.8147 for UE. We saw no additional need to test the white noise hypothesis on composited data.

There appears to be no statistical gain in the efficiency of CPS estimates due to improved stratification; i.e., annual sampling seems to have no effect on the efficiency of CPS monthly estimates.

7.3 Theoretical Issues of MAF Adds and Unduplication Across Sample Selections: Conditional Probabilities of Selection under Annual Sampling

We discuss the conditional probabilities of selection under the following assumptions:

- We may change the probabilities of selection from year-to-year, such as for unduplication with other surveys, sample cuts, expansions, or to maintain a fixed sample size when the size of the frame has changed.
- A HU may be selected at most once within any five-year period.

7.3.1 Sample space

Our approach to the problem is axiomatic, in which joint probabilities are defined across an arbitrary number of years. We distinguish between the sampling universe (sample space) and the sampling frames to allow us to derive conditional and joint probabilities of selection for multiple points in time using fundamental rules of probabilities. We assume the unconditional probabilities are known beforehand, and make no further assumptions about them.

For a given HU, our sample space is all combinations of possible outcomes of sampling across n years. For example, if *H = "in sample," T = "not in sample,"* and *n=10*, one possible element would be *HHTTHHTTHH.* Many of these combinations will have probability zero, such as those that represent *"in sample"* in year one when the HU was first introduced to the sampling frame at a later year, or combinations that violate the restriction that HUs can only be selected once in any five-year period. Still, those combinations will be included in our sample space. An example of an event for this probability space would be *"in sample at year 2"* which, for *n=10,* would be all combinations of *H*s and *T*s of length 10 with an *H* in the second position. As was already noted, the unconditional probabilities are predetermined and are not necessarily equal to the count of the number of elements in an event divided by the number of elements in the sample space.

7.3.2 Conditional probabilities of selection

This section shows the conditional probabilities of selection for all HUs eligible to be sampled in a given year. HUs that are ineligible for sample due to having been recently sampled will have conditional probability zero. Let A_i be the event that a given HU is in sample at year $i=1, 2, \ldots, n$.

A HU may be selected once, at most:
Up to year five, there is zero probability that a HU will be in sample more than once. Define the zero'th year outcome to be the null set, so we can express the conditional probability for year one (conditioned on A_0^c) and it will be equal to the unconditional probability.

$$\text{Let } p_i = P(A_i) \text{ and } \tilde{p}_i = P\left(A_i \mid \bigcap_{j=1,\ldots,i-1} A_j^c \right), \text{ for } i=2,3,4,5, \text{ with } \tilde{p}_1 = p_1 .$$

$$\text{Then } \tilde{p}_i = \frac{p_i}{1-\sum_{j=1}^{i-1} p_j} \text{ , for } i=2,3,4,5.$$

After four years, a HU may re-enter sample:
If a HU has not been in sample for four consecutive years, then we will allow it to enter sample, regardless of whether it had ever been selected.

Loosening the restriction on re-entering sample gives us the freedom to choose a constraint other than the zero-probability constraint used in the previous case. There may be any number of choices that will lead to a valid solution, but one that seems sensible is to make the selection probability dependent only on the previous four years' selections. In fact, this led to a solution with a form that meshes nicely with the previous result.

We will alter the definition of $\tilde{p}_i$ so that it depends only on the previous four years' samples. That is,

$$\tilde{p}_i = P\left(A_i \mid \bigcap_{j=\max(1,i-4),\ldots,i-1} A_j^c \right) \text{ and it follows that } \tilde{p}_i = \frac{p_i}{1-\sum_{j=i-4}^{i-1} p_j} \text{, for } i=5,6,\ldots,n.$$

7.3.3 Unduplication with other surveys and the use of subframes

We have discussed probabilities of selection that condition on the outcomes in prior years for a given HU. We may also consider conditioning on additional events. Redefining $\tilde{p}_i$ and p_i so that they condition on a random trial R leads to results in the same form as those shown earlier.

One option for annual sampling that has been discussed is through the use of subframes. This involves a trial R that distributes sample into five random groups such that we alternate which group we select sample from year-to-year. In this case, the p_i's will be nonzero only once every five years, so the results will reduce to $\tilde{p}_i = p_i$. (See Rottach (2009) for further details.)

8. Summary and Continuing CPS Research

The simulation of the month-to-month and the year-to-year correlations are statistically different under annual sampling but the changes in correlation seem to be small enough that they will not drastically change the variance on the key change estimates in the CPS. Although annual sampling would require a 5 percent increase in sample from the current 60,000 Basic CPS HUs to maintain the requirement on estimates of changes in UE, the CV reliability for the UE level for the nation that we should control to, given a 6 percent UE rate and given the requirement that a difference of 0.2 percentage points in the UE rate for two consecutive months be statistically significant at the 0.10 level, would decrease approximately to 1.85 percent, which is within rounding of the current reliability requirement of 1.9 percent. In terms of the fifth survey requirement listed in Section 4, simulation results showed that annual sampling decreases the annual average variance; thus, this requirement is satisfied since it only affects the distribution of sample among the states and not the overall sample size.

Based on the available data, there does not exist a trend in design effects, which points to the fact that we are not seeing a decline in efficiency caused by using 2000 census sort variables as we get farther away from the 2000 Decennial Census. Lastly, the theoretical issues of the MAF adds and unduplication across sample selections can be resolved in a manner which is easy to implement.

If the CPS were to implement annual sampling, then there would be a slight loss in efficiency in that it could require a 5 percent increase in sample to maintain the requirement on estimates of changes in UE. As stated before, this 5 percent increase could be reduced if the additional sample could be distributed more effectively across the states. This will be a priority of our continuing research. Accepting that, then annual sampling could be done, if the costs were not prohibitive.

The analysis provided will continue to be extended and refined. Perhaps we will examine more frequent sampling than annually, such as every two years; we would initiate our efforts into this research via design effects. Along these same lines, we may consider selecting a new sample for each sample designation. Then, you are not really looking at a 100 percent new sample each year. The new sample is still phased in, so the benefits in terms of adjusted sample sizes and more up-to-date will only be realized gradually. Initially we don't think there would be major differences in the pros and cons from those of annual sampling, but it seems like it might be less complicated, and easier to explain.

Acknowledgements

The authors thank Samson A. Adeshiyan, CPS Branch Chief of the Demographic Statistical Methods Division, for his guidance and statistical support. Also, our thanks are extended to Lindsay McMillan for his assistance in the running and analysis of design effects, and Michael Ikeda for his statistical review of this paper.

Supporting Material

Bureau of Labor Statistics and U.S. Census Bureau (2009). *2009 IAA: Interagency Agreement Between the Bureau of Labor Statistics (BLS) and the U.S. Department of Commerce, U.S. Census Bureau*.

Fay, R.E. (1984). "Some Properties of Estimators of Variance Based on Replication Methods." Proceedings of the Section on Survey Research Methods, American Statistical Association. Alexandria, VA. pp. 495-500.

Fay, R.E. (1989). "Theory and Application of Replicate Weighting for Variance Calculations." 1989 Proceedings of the Section on Survey Research Methods, American Statistical Association. Alexandria, VA. pp. 212-217.

Fay, R.E. and Train, G. (1995). "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties." Proceedings of the Government Statistics Section, American Statistical Association. Alexandria, VA. pp. 154-159.

Fisher, R.A. (1915). "Frequency Distribution of the Values of the Correlation Coefficient in Samples of an Indefinitely Large Population." Biometrika 10: pp. 507–521.

Fisher, R.A. (1921). "On the `Probable Error' of a Coefficient of Correlation Deduced from a Small Sample." Metron 1: pp. 3–32.

Memorandum from 2000 Redesign Work Group on Future MAF Sampling for New Construction, New Surveys, and Expansions (WG 4.5). "Work Plan for WG 4.5" (Doc. # 4.5-C-1), Attachment A Definitions. September 21, 1999.

Memorandum from 2010 Sample Redesign Demographic Surveys Sample Design Work Group 3.0. "Charter for the Within-PSU MAF Sampling Team, WG 3.8 (Doc. #2010-3.8-C-1, Version 0.3)." August 26, 2008.

Memorandum from Lawrence S. Cahoon for James M. Lewis. "Reassigning Within-Hit Numbers in the Unit Frame for the 2000 Sample Design for CPS/SCHIP and NCVS (Doc. #4.8-S-5)." February 13, 2004.

Memorandum from Patrick Flanagan. "Bureau of Labor Statistics (BLS) Unduplication Proposal: Points to Consider." June 16, 2008.

Minutes from WG 3.0 Meeting on April 15, 2008. "Attachment B: Three Alternatives for 2010 DSSR Updating and Unduplication (Doc. #2010-3.0-M-49)." May 23, 2008.

Minutes from the Bureau of Labor Statistics (BLS)/Census Bureau Meeting on Redesign Research of 27 June 2008, minute-taker Patrick Flanagan. June 27, 2008.

Rottach, R. (2009). "Conditional Probabilities of Selection With Annual Sampling." Census Bureau Memorandum for Documentation from Reid Rottach.

SAS Institute Inc. (2004) SAS/ETS© 9.1 *User's Guide*. Cary, NC: pp. 365-480.

U.S. Census Bureau (2006). *Current Population Survey: Design and Methodology*, Technical Paper 66. October 2006.

U.S. Census Bureau (2007). X-12-ARIMA Reference Manual, Version 0.3. http://www.census.gov/ts/x12a/v03/x12adocv03.pdf.

[i] In past designs, the current surveys independently sampled without replacement from a skeleton universe of theoretical sample measures within each Basic PSU Component (BPC). The resulting layout of sample measures taken by each survey within the BPC is called the skeleton sample.

[ii] For purposes of this paper, 'the MAF' refers to the current MAF extract, updated by the most recent DSF, the Community Address Update System (CAUS) of the Decennial Statistical Studies Division of the Census Bureau, and all other city- and non-city-style update/correction sources.

[iii] The area frame collects old and new construction in locations not covered by a permit office.

[iv] A BPC is the intersection of all or part of all surveys' sample stratification PSU definitions.

[v] Analogous processes are used to select sample in the other three frames.

[vi] Although ACS data will be used in this CPS research, annual sampling allows the option of using Decennial or ACS data, with administrative records, choosing whichever provides the most up-to-date data for the second-stage sample selection.

[vii] Additions to the MAF from new construction or other additions.

[viii] By 'without replacement' we mean a HU will not be selected in a future annual sample if it had been selected in a prior one.

[ix] Depending on our approach, we may have independence, or just approximately so.

[x] X-12-ARIMA is the seasonal adjustment software written, developed, and maintained at the U.S. Census Bureau. The current official version of X-12-ARIMA is Version 0.3 released May 2007. It is used for all official seasonal adjustments at the U.S. Census Bureau.

[xi] The ARIMA procedure that is available in SAS provides a set of tools for univariate time series model identification, parameter estimation, and forecasting. It offers flexibility in the kinds of ARIMA models that can be analyzed, using the AutoRegressive Integrated Moving-Average (ARIMA) model. An ARIMA model predicts a value in a response time series as a linear combination of its own past values, past errors, and current and past values of other time series.

[xii] The negative correlations for UE estimated at higher lags, and the large corresponding differences between the simulated annual sampling and the current approaches would not be expected. They are likely due to random variation and the string of negative estimates may reflect the fact that the tests are not independent.