BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION
RESEARCH REPORT SERIES
No. RR-91/10

MULTIPLE WORKLOADS PER STRATUM DESIGNS

by

Lynn Weidman
Lawrence R. Ernst
U.S. Bureau of the Census
Statistical Research Division
Washington, D.C. 20233

Report issued: March 15, 1993 (Revised)

Abstract: This paper introduces an approach to expanding a stratified sample design, D_1, with one primary sampling unit (PSU) selected per stratum to a larger design, D_2. Define a workload (WL) to be the sample size in a given stratum in D_1. This three-stage approach selects the number of WLs for each stratum, the PSUs to receive the additional WLs in each stratum, and the ultimate sampling units. Procedures requiring the consideration of several cases are given for selecting PSUs in the key second stage, satisfying the following conditions when a stratum in D_2 is to have s=2 or 3 WLs: (i) the expected number of WLs in a PSU is s times the probability that it was selected to get the single WL in D_1; and (ii) the actual number of WLs assigned is within one of the expected number. These conditions are a generalization of probability proportional to size, without replacement sampling. A decomposition of the variance into components for the three selection stages is derived. An application of this approach to a proposed, but since cancelled, expansion of the Current Population Survey is also presented.

Key words: Stratified sample design; PSU selection; workload; variance decomposition.

1. Introduction

Consider the situation where there is a survey currently in operation having a stratified sample design, D_1, with one primary sampling unit (PSU) selected per stratum. At some time it is necessary to make a substantial increase in the sample size of this survey to meet new variance criteria, while retaining all the originally designated PSUs. In this paper we describe one approach to expanding the sample which involves a new methodology.

The basic method that we use is to select additional PSUs for the expanded design, D_2, from the D_1 strata and to then select a sample in each added PSU which ideally would be of the same size as the D_1 sample in that stratum. (We call this sample size a *workload* (WL).) We would prefer that no PSU be selected more than once for the D_2 design, since without replacement sampling is generally more efficient than with replacement sampling. However, this goal cannot always be met since we require that the expected number of times that a PSU is selected be proportional to the size of the PSU. To illustrate this point, suppose that the size of D_2 is twice that of D_1. One approach for this case is to select a second, distinct PSU from each stratum using the Brewer-Durbin procedure (Cochran, 1977, pp 261-263). However, if one PSU constitutes more than half the size of a stratum then this approach must fail, since the expected number of times that this PSU would be selected for the D_2 design is twice its D_1 selection probability, and hence greater than one. Consequently, this PSU must have a positive probability of being selected twice in the D_2 design. It is still possible, though, to minimize the variability in the number of times this large PSU is selected. While an independent selection of the two PSUs from the stratum to be in the sample could result in the large PSU being selected either 0, 1 or 2 times, the procedure to be presented will insure that this PSU is selected at least once.

In general, in this paper the requirements imposed on the procedure for sample PSU selection for D_2 are the appropriate generalization of probability proportional to size, without replacement sampling to the case when s WLs are to be selected in a stratum without any restrictions on the relative sizes of the PSUs in the stratum. They are: (*i*) the expected number of WLs in a PSU is s times the probability that it was selected to get the single WL in D_1; and (*ii*) the actual number of WLs assigned is within one of the expected number. In addition, we require that each

D_1 sample PSU be a D_2 sample PSU. We provide complete details on how this can be accomplished for s=2 and s=3, which requires consideration of a number of different cases.

This work was motivated by a formerly planned expansion of the Current Population Survey (CPS) that was to be selected in two phases. Phase 1 would be a redesign of the present CPS that must meet monthly variance requirements on estimates of number of persons unemployed for the nation, the eleven largest states, New York City and Los Angeles. At a later date, phase 2 would select additional sample to meet similar monthly requirements for the remaining 39 states and the District of Columbia. The approach presented in this paper was one of several options investigated for this two-phase sampling. Each of the other options has at least one of the following drawbacks: the phase 2 sample PSUs must be selected simultaneously with the phase 1 PSUs; some phase 1 sample PSUs are dropped from sample in phase 2; or small PSUs can receive multiple WLs in phase 2. The approach in this paper avoids all of these drawbacks. It has the advantage that it is based solely on the stratification and initial selection probabilities used for phase 1, and phase 2 principally involves selecting PSUs from these strata to receive the additional sample. Although the CPS application motivated this work, there are potential applications to other sample expansion problems.

Section 2 describes the entire procedure for expanding from D_1 to D_2, which includes two other stages of sampling in addition to selection of PSUs. Section 3 details the procedures for selecting PSUs to receive the additional WLs, which is the only sampling stage for which the selection methodology is not routine. Variance formulae are derived in Section 4. Finally, Section 5 presents as an example the variances for the CPS application and compares them with other options investigated at the Census Bureau.

2. Expanding an Existing Design

The presentation in Sections 2-4 considers the expansion from D_1 to D_2 for noncertainty PSUs only. For certainty PSUs, the expansion is obtained by simply selecting an appropriate number

of ultimate sampling units (USUs) to supplement the D_1 sample.

The expansion from D_1 to D_2 using the multiple WLs approach requires the following three stages of selection to be carried out. First, the total number of WLs are allocated among the strata. Then the WLs in each stratum are allocated among its PSUs. Finally, USUs comprising the WLs are selected within the designated PSUs. We proceed to describe each of these three stages, with the details on the second stage, which is the focus of this paper, postponed until Section 3. We introduce necessary notation at the beginning of the description of each stage.

Let n_1 denote the number of D_1 sample USUs and n_2 the desired number of D_2 sample USUs. Let I denote the number of strata in the designs. Then ideally, the number of D_2 WLs, denoted T, would be n_2I/n_1, since in each stratum this would result in the same D_1 and D_2 WL sizes. Since n_2I/n_1 is generally not an integer, we let $T = \lceil n_2I/n_1 \rceil$, resulting in a generally slightly smaller WL size in each stratum for D_2 than for D_1. (For any number x, let $\lfloor x \rfloor$ and $\lceil x \rceil$ denote the greatest integer not exceeding x and the smallest integer not less than x, respectively.) Define $R=T/I$, so at the first stage of sampling each stratum is assigned either $\lfloor R \rfloor$ or $\lfloor R \rfloor+1$ WLs, with a simple random sample of $T-I\lfloor R \rfloor$ strata receiving $\lfloor R \rfloor+1$ WLs. (This allocation limits the variation in the number of WLs each stratum can receive and hence reduces the between strata component of variance.) We let s_i denote the number of WLs assigned to the i-th stratum by the first stage of sampling. Note that

$$E(s_i) = R, \quad i=1,...,I. \tag{2.1}$$

Let J_i, $i=1,..,I$, denote the number of PSUs in the i-th stratum; let p_{ij} be the probability of selection of the j-th PSU in the i-th stratum in the D_1 design; and let w_{ij} be the number of WLs assigned to this PSU in the D_2 design. At the second stage of sampling for the D_2 design, the s_i WLs in the i-th stratum are allocated among the PSUs in that stratum in such a manner that the following conditions are satisfied:

$$w_{ij} \geq 1 \text{ for each PSU } ij \text{ in sample for the } D_1 \text{ design;} \tag{2.2}$$

$$w_{ij} = \lfloor p_{ij} s_i \rfloor \text{ or } w_{ij} = \lfloor p_{ij} s_i \rfloor + 1; \tag{2.3}$$

$$P(w_{ij} = \lfloor p_{ij} s_i \rfloor + 1) = p_{ij} s_i - \lfloor p_{ij} s_i \rfloor. \tag{2.4}$$

Note that it follows from (2.3) and (2.4) that

$$E(w_{ij} \mid s_i) = p_{ij} s_i, \tag{2.5}$$

while (2.1) and (2.5) yield the unconditional expectation

$$E(w_{ij}) = p_{ij} R. \tag{2.6}$$

If s_i=1, then the D_2 sample PSU for stratum i is simply the D_1 sample PSU by (2.2). For s_i=2 and s_i=3, a procedure for allocating WLs that satisfies (2.2) - (2.4) is presented in Section 3. It would be possible to extend this procedure for values of s_i>3 using similar approaches, although with increasing complexity as s_i increases.

Let N, N_i, and N_{ij} be the number of USUs in the total population, the i-th stratum, and the ij-th PSU, respectively. We assume that p_{ij}=N_{ij}/N_i and that the USUs within sample PSUs are selected in a manner such that each USU in the population has the same probability of selection, that is n_2/N. Consequently, the WL size for D_2 in the i-th stratum, denoted m_i, is

$$m_i = \frac{n_2 N_i}{N R}. \tag{2.7}$$

Within each PSU ij in the i-th stratum for which w_{ij}>0, $w_{ij}m_i$ USUs are selected with equal probability. Irrespective of how this is done, an unbiased estimator $\hat{y}$ of total for a characteristic y is given by

$$\hat{y} = \frac{N}{n_2} \sum y_{ijk}, \tag{2.8}$$

where y_{ijk} is the total value for characteristic y for a sample USU from PSU ij, and the summation is over all sample USUs. Some y_{ijk}'s may possibly appear more than once in this summation if with replacement sampling is used at the final stage. A formula for the variance of $\hat{y}$ under our three-stage sampling approach is presented in Section 4.

For the three-stage sampling procedure just outlined, the expected number of D_2 sample USUs is n_2, but the actual number selected is a random variable that depends on the first stage of sampling. This is because, by (2.7), the WL size is not the same for each stratum. An alternative three-stage procedure for which the number of sample USUs would always be n_2 would begin by selecting the number of WLs, s_i, assigned to the i-th stratum so that $E(s_i) = TN_i/N$, that is proportional to size. For every sample, s_i would also be within 1 of $E(s_i)$. The second and third stages would be selected as described above, except now the WL size in each stratum would be n_2/T. This alternative approach will not be discussed further in the paper.

3. **Selecting PSUs to Receive Additional Workloads**

To simplify notation, we drop the subscript for stratum in this section only. To complete the specifications for the sampling procedure it remains only to state procedures for selecting s-1 additional WLs, s=2,3, in a stratum for the D_2 design when a single PSU, j, has been selected for the D_1 sample with probability p_j, with the s WLs satisfying (2.3), (2.4).

The procedures to be detailed will also satisfy the following additional condition:

Each selection ordering for a set of s WLs is equally likely. (3.1)

We order the PSUs in the stratum to satisfy $p_1 \geq p_2 \geq \ldots \geq p_J$ and introduce the following additional notation, with the first WL being the D_1 WL for the stratum.

$P(k|j)$ = $P(PSU\ k$ gets second $WL|PSU\ j$ got first WL),

$P(k\ell|j)$ = $P(PSU\ k$ and $PSU\ \ell$ get second and third WLs in any order$|PSU\ j$ got first WL),

$P(jk)$ = $P(PSU\ j$ and $PSU\ k$ each get 1 WL in any order),

$P(jk\ell)$ = $P(PSU\ j$, $PSU\ k$ and $PSU\ \ell$ each get 1 WL in any order).

There are several cases to consider which depend on the values of $\lfloor sp_1 \rfloor$ and $\lfloor sp_2 \rfloor$. Two of the cases are direct applications of the methods of Brewer-Durbin (Cochran 1977) for two-PSU-per-stratum designs or Sampford (1967) for designs of three or more PSUs per stratum. For the other cases with s=3, conditional probabilities $P(k\ell|j)$ are obtained as follows. Joint probabilities, $P(jk\ell)$, are presented which satisfy (2.3) and (2.4). From these joint probabilities, conditional probabilities, $P(k\ell|j)$, which satisfy (3.1) can then be obtained by dividing $P(jk\ell)$ by p_j and multiplying the result by the proportion of the selections of j,k,ℓ for which j is selected first. This proportion is 1/3 if j is distinct from both k and ℓ, 2/3 if j equals exactly one of k,ℓ, and 1 if $j=k=\ell$. Conditional probabilities $P(k|j)$ for s=2 are obtained similarly from $P(jk)$.

A. s=2.

1. $p_1 \geq 1/2$. Let $P(11) = 2p_1 - 1$, $P(1j) = 2p_j$, $j \neq 1$, which satisfies (2.4) for j=1, and $j \neq 1$, respectively. Then,

   $P(1|1) = 2 - 1/p_1$,

   $P(j|1) = p_j/p_1$, $j \neq 1$,

   $P(1|j) = 1$, $j \neq 1$.

2. $p_1 < 1/2$. Use the Brewer-Durbin procedure, that is

$$P(k|j) = p_k\left[\frac{1}{1-2p_j}+\frac{1}{1-2p_k}\right] \Big/ \left[1 + \sum_{\ell=1}^{J}\frac{p_\ell}{1-2p_\ell}\right]$$

B. s=3.

1. $p_1 \geq 2/3$. Let $P(111) = (3p_1-2)$, $P(11j) = 3p_j$, $j\neq 1$, which satisfies (2.4) for j=1, and $j\neq 1$, respectively. Then,

$P(11|1) = (3p_1-2)/p_1$,

$P(1j|1) = 2p_j/p_1$, $j\neq 1$,

$P(11|j) = 1$, $j\neq 1$.

2. $p_1,p_2 \geq 1/3$. Let $P(112) = 3p_1-1$, $P(122) = 3p_2-1$, and $P(12j) = 3p_j$, $j\neq 1,2$, which satisfies (2.4) for j=1, j=2, and j>2, respectively. Then for j,k,ℓ with $\{j,k\}=\{1,2\}$, $\ell\neq 1,2$,

$P(kk|j) = (3p_k-1)/(3p_j)$,

$P(jk|j) = 2(3p_j-1)/(3p_j)$,

$P(k\ell|j) = p_\ell/p_j$,

$P(12|\ell) = 1$.

3. $1/3 \leq p_1 < 2/3$, $p_2 < 1/3$. Let $p_1' = (3p_1-1)/2$, $p_j' = 3p_j/2$, $j\neq 1$,

$$P(1jk) = D(j,k), \quad j\neq k \ (j \text{ or } k \text{ can be } 1), \tag{3.2}$$

where

$$D(j,k) = 2p_j' p_k'\left[\frac{1}{1-2p_j'}+\frac{1}{1-2p_k'}\right]\Big/\left[1+\sum_{\ell=1}^{J}\frac{p_\ell'}{1-2p_\ell'}\right].$$

Then, since the $D(j,k)$ are Brewer-Durbin joint probabilities, it follows that

$$\sum_{k\neq 1} P(11k) = \sum_{k\neq 1} D(1k) = 2p_1' = 3p_1-1,$$

and hence (2.4) is satisfied for $j=1$; similarly for $j\neq 1$,

$$\sum_{\substack{k \\ k\neq j}} P(1jk) = \sum_{\substack{k \\ k\neq j}} D(jk) = 2p_j' = 3p_j,$$

and hence (2.4) is satisfied for $j\neq 1$. Then, by (3.2),

$$P(1j|1) = 2D(1j)/(3p_1), \quad j\neq 1$$

$$P(jk|1) = D(jk)/(3p_1), \quad j,k\neq 1,\ j\neq k$$

$$P(1k|j) = D(jk)/(3p_j), \quad j\neq 1,\ k\neq j \ (k \text{ can equal } 1).$$

4. $p_1<1/3$. Use Sampford's procedure, that is

$$P(k\ell|j) = \frac{K_3 p_k p_\ell}{3}\left[\frac{1}{(1-3p_j)(1-3p_k)} + \frac{1}{(1-3p_j)(1-3p_\ell)} + \frac{1}{(1-3p_k)(1-3p_\ell)}\right],$$

where

$$K_3 = \left[\frac{L_2}{3} + \frac{2L_1+1}{9}\right]^{-1}, \ L_1 = \sum_{\ell=1}^{J} \lambda_\ell, \ L_2 = \sum_{\substack{j=1 \\ j\neq k}}^{J}\sum_{k=1}^{J} \lambda_j\lambda_k, \ \lambda_\ell = \frac{p_\ell}{1-3p_\ell}.$$

4. Variance Decomposition

We proceed to develop a formula for $V(\hat{y})$, the variance of the estimator $\hat{y}$ of (2.8). There will be three terms in $V(\hat{y})$, reflecting the three stages of sampling.

We first obtain an alternative expression for $\hat{y}$. Let

$$\hat{y}_{ij} = N_{ij}\sum_{k} y_{ijk}/(w_{ij}m_i) \qquad \text{if } w_{ij}>0, \tag{4.1}$$

$$= 0 \qquad \text{if } w_{ij}=0.$$

Given $w_{ij}>0$, $\hat{y}_{ij}$ is an unbiased estimator of the total, denoted y_{ij}, for PSU ij. Then combining (2.7), (2.8), and (4.1) we obtain

$$\hat{y} = \frac{N}{n_2}\sum_{ijk} y_{ijk} = \frac{N}{n_2}\sum_{i=1}^{I} m_i \sum_{j=1}^{J_i} \frac{w_{ij}\hat{y}_{ij}}{N_{ij}} \tag{4.2}$$

$$= \frac{1}{R}\sum_{i=1}^{I} N_i \sum_{j=1}^{J_i} \frac{w_{ij}\hat{y}_{ij}}{N_{ij}} = \sum_{i=1}^{I}\sum_{j=1}^{J_i} a_{ij}\hat{y}_{ij},$$

where

$$a_{ij} = \frac{w_{ij}N_i}{R\,N_{ij}} = \frac{w_{ij}}{Rp_{ij}} \tag{4.3}$$

is a random variable whose value for each sample is determined by the first two sampling stages.

The variance of $\hat{y}$ can then be written in the form

$$V(\hat{y}) = V_1E_2E_3\left(\sum_{i=1}^{I}\sum_{j=1}^{J_i} a_{ij}\hat{y}_{ij}\right) + E_1V_2E_3\left(\sum_{i=1}^{I}\sum_{j=1}^{J_i} a_{ij}\hat{y}_{ij}\right) + E_1E_2V_3\left(\sum_{i=1}^{I}\sum_{j=1}^{J_i} a_{ij}\hat{y}_{ij}\right) \tag{4.4}$$

The subscripts on the expectations denote the three stages of sampling. Since whenever $w_{ij}>0$, $E_3(\hat{y}_{ij}) = y_{ij}$, then

$$V(\hat{y}) = V_1\left[\sum_{i=1}^{I}\sum_{j=1}^{J_i} E_2(a_{ij})y_{ij}\right] + E_1\left[\sum_{i=1}^{I} V_2\left(\sum_{j=1}^{J_i} a_{ij}y_{ij}\right)\right] + E_1\left[\sum_{i=1}^{I}\sum_{j=1}^{J_i} E_2\left\{a_{ij}^2 V_3(\vartheta_{ij})\right\}\right]. \qquad (4.5)$$

These three terms are, respectively, the between strata, between PSUs within strata, and within PSUs components of variance.

4.1 Between Strata Variance

Let $y_i = \sum_{j=1}^{J_i} y_{ij}$, the characteristic total for stratum i. Let S denote the set of strata i for which $s_i = \lfloor R \rfloor + 1$ and let $Z = T - I\lfloor R \rfloor$, that is the number of such strata. If $Z=0$, then the between strata component of variance is 0. Otherwise, by (2.5), (4.3), and the fact that $\lfloor R \rfloor\, y_i$ is a constant, we have

$$V_1\left[\sum_{i=1}^{I}\sum_{j=1}^{J_i} E_2(a_{ij})y_{ij}\right] = V_1\left(\sum_{i=1}^{I}\sum_{j=1}^{J_i} \frac{s_i}{R} y_{ij}\right) = \frac{1}{R^2} V_1\left(\sum_{i=1}^{I} s_i y_i\right)$$

$$= \frac{1}{R^2} V_1\left(\sum_{i=1}^{I} (s_i - \lfloor R \rfloor)\, y_i\right) = \left(\frac{Z}{R}\right)^2 V_1\left(\sum_{i\in S} \frac{y_i}{Z}\right). \qquad (4.6)$$

Since the set S is a simple random sample of Z out of I strata, (4.6) is $(Z/R)^2$ times the variance of the mean from a simple random sample. Using Theorem 2.2 of Cochran (1977), we obtain that (4.6) reduces to

$$\frac{(I-Z)Z}{R^2 I(I-1)}\left[\sum_{i=1}^{I}\left(y_i-\frac{y}{I}\right)^2\right]. \tag{4.7}$$

4.2 Between PSUs Within Strata Variance

Let $w'_{ij} = w_{ij} - \lfloor s_i\, p_{ij} \rfloor$. To evaluate the second bracketed term in (4.5), we first expand

$$V_2\left(\sum_{j=1}^{J_i} a_{ij}\, y_{ij}\right) = \frac{1}{R^2}\, V_2\left(\sum_{j=1}^{J_i} w_{ij}\, \frac{y_{ij}}{p_{ij}}\right) = \frac{1}{R^2}\, V_2\left(\sum_{j=1}^{J_i} w'_{ij}\, \frac{y_{ij}}{p_{ij}}\right)$$

$$= \frac{1}{R^2}\left(\sum_{j=1}^{J_i} E_2[(w'_{ij})^2]\, \frac{y_{ij}^2}{p_{ij}^2} + \sum_{\substack{j=1 \\ j\neq k}}^{J_i}\sum_{k=1}^{J_i} E_2(w'_{ij}\, w'_{ik})\, \frac{y_{ij}y_{ik}}{p_{ij}p_{ik}} - \left[\sum_{j=1}^{J_i} E_2(w'_{ij})\, \frac{y_{ij}}{p_{ij}}\right]^2\right), \tag{4.8}$$

where the substitution of w'_{ij} for w_{ij} is justified by the fact that conditioned on s_i, these two random variables differ by a constant.

Now, since w'_{ij} is a 0-1 variable, it follows that

$$E_2[(w'_{ij})^2] = E_2(w'_{ij}) = E_2(w_{ij}) - \lfloor s_i p_{ij} \rfloor = s_i p_{ij} - \lfloor s_i p_{ij} \rfloor. \tag{4.9}$$

Consequently, it remains only to evaluate $E_2(w'_{ij}\, w'_{ik})$ in order to obtain a complete expression for (4.8).

Let

$$Q_{ijks_i} = P(w_{ij} = \lfloor s_i p_{ij} \rfloor + 1 \text{ and } w_{ik} = \lfloor s_i p_{ik} \rfloor + 1), \quad j \neq k. \tag{4.10}$$

Then $E_2(w'_{ij}\, w'_{ik}) = Q_{ijks_i}$. If $s_i = 1$, then clearly $Q_{ijkl} = 0$ for all j,k. In addition, $Q_{ijks_i} = 0$ for many of the cases listed in Section 3 for which $s_i = 2$ or $s_i = 3$. For example, $Q_{ijk2} = 0$ for all j,k for Case A.1. This is because $\lfloor 2p_{i1} \rfloor + 1 = 2$, $\lfloor 2p_{ij} \rfloor + 1 = 1$, $j \neq 1$, and hence there is no pair j,k, $j \neq k$, for which $w_{ij} = \lfloor 2p_{ij} \rfloor + 1$ and $w_{ik} = \lfloor 2p_{ik} \rfloor + 1$, since PSU 1 must have at least 1 of the 2 WLs. The only situations for which $Q_{ijks_i} \neq 0$ with $s_i = 2$ or $s_i = 3$ are presented below. To simplify notation, the stratum subscript i is dropped from Q_{ijks_i}, p_{ij}, and J_i.

Case A.2. $$Q_{jk2} = P(jk) = 2p_j p_k \left[\frac{1}{1-2p_j} + \frac{1}{1-2p_k} \right] \Big/ \left[1 + \sum_{j=1}^{J} \frac{p_k}{1-2p_k} \right]$$

(Cochran 1977, p. 262).

Case B.3. By (3.2), $Q_{1k3} = P(11k) = D(1,k)$, $k \neq 1$.

Case B.4. $$Q_{jk3} = P(jk) = K_3 \lambda_j \lambda_k \left[(2-3p_j-3p_k) \left(\sum_{l=1}^{J} \lambda_l - \lambda_j - \lambda_k \right) + 1 - p_j - p_k \right],$$

where K_3 and λ_j are as defined in Section 3 (Sampford 1967).

Finally, let

$$f(\lfloor R \rfloor) = 1 + \lfloor R \rfloor - R, \qquad f(\lfloor R \rfloor + 1) = R - \lfloor R \rfloor, \tag{4.11}$$

the probabilities that a stratum receives $\lfloor R \rfloor$ and $\lfloor R \rfloor + 1$ WLs, respectively. Then, combining (4.8)-(4.11), we conclude

$$E_1\left[\sum_{i=1}^{I} V_2\left(\sum_{j=1}^{J_i} a_{ij} y_{ij}\right)\right] = \sum_{i=1}^{I} \sum_{\ell=\lfloor R \rfloor}^{\lfloor R \rfloor + 1} V_2\left[\left(\sum_{j=1}^{J_i} a_{ij} y_{ij}\right) \mid s_i = \ell\right] f(\ell)$$

$$= \frac{1}{R^2} \sum_{i=1}^{I} \sum_{\ell=\lfloor R \rfloor}^{\lfloor R \rfloor + 1} \left(\sum_{j=1}^{J_i} (\ell p_{ij} - \lfloor \ell p_{ij} \rfloor) \frac{y_{ij}^2}{p_{ij}^2} + \sum_{\substack{j=1 \\ j \neq k}}^{J_i} \sum_{k=1}^{J_i} Q_{ijk\ell} \frac{y_{ij} y_{ik}}{p_{ij} p_{ik}} \right.$$

$$\left. - \left[\sum_{j=1}^{J_i} (\ell p_{ij} - \lfloor \ell p_{ij} \rfloor) \frac{y_{ij}}{p_{ij}} \right]^2 \right) f(\ell). \tag{4.12}$$

4.3. Within-PSUs Variance

Finally we evaluate the third bracketed term in (4.5) under the assumption that all sample WLs within a PSU are selected by simple random sampling, either with replacement for large N_{ij} or without replacement with a negligible finite population correction factor. Appropriate modifications are necessary for other within PSU selection procedures.

From (4.1) it follows that if $w_{ij}>0$, then

$$V_3(\hat{y}_{ij}) \doteq \frac{N_{ij}^2 S_{3ij}^2}{w_{ij} m_i}, \tag{4.13}$$

where S_{3ij}^2 is the population variance of y in PSU ij.

Furthermore, by (2.6), (2.7) and (4.3),

$$E_1E_2\left(\frac{a_{ij}^2N_{ij}}{w_{ij}m_i}\right) = \frac{E_1E_2(w_{ij})N_{ij}}{R^2m_ip_{ij}^2} = \frac{N_{ij}}{Rm_ip_{ij}} = \frac{N_i}{Rm_i} = \frac{N}{n_2}, \tag{4.14}$$

which we then combine with (4.13) to obtain

$$E_1\left[\sum_{i=1}^{I}\sum_{j=1}^{J_i} E_2\left(a_{ij}^2V_3(\vartheta_{ij})\right)\right] = \frac{N}{n_2}\sum_{i=1}^{I}\sum_{j=1}^{J_i} N_{ij}S_{3ij}^2. \tag{4.15}$$

Finally, if the S_{3ij}^2 are the same for all ij, with common value denoted S_3^2, then by summing (4.15) over all ij we obtain that the within PSUs variance is approximately $N^2S_3^2/n_2$. This is approximately the sampling variance for the standard estimator of population total from a simple random sample with replacement, for a sample of size n_2 selected from a population of size N, for N large, with variance S_3^2. Similar assumptions lead to the same approximate within PSUs variance for the other options investigated for the two phase sampling application, a result which will be used in some of the comparisons in the next section.

5. Comparison of Methods for the CPS Expansion

In this section, variances for the multiple WLs per stratum method are compared to variances for three other methods for selecting the D_2 sample for the formerly planned CPS expansion,

discussed in Section 1. The other three methods are the independent sample, the independent supplement (both described in Chandhok, Weinstein and Gunlicks (1990)), and controlled selection (Ernst, 1990). The independent sample method selects the D_2 sample PSUs from an optimal D_2 stratification independently of the D_1 sample PSUs. The controlled selection method simultaneously selects sample PSUs for D_1 and D_2 from optimal stratifications for these two designs, while insuring, unlike the independent sample method, that the D_1 sample PSUs are a subset of the D_2 sample PSUs. The independent supplement method includes all D_1 sample PSUs in D_2 and selects additional sample PSUs for inclusion in D_2 independently from a second, supplemental stratification.

In Table 1, the ratio of variances for controlled selection, independent supplement, and multiple WLs methods, to the independent sample method are presented. These total variances include the within PSUs component from both certainty and noncertainty PSUs. For all four methods, 1980 census data were used to obtain the stratifications, since 1990 census data were unavailable at the time these computations were done. The variances were computed using 1970 data to simulate a 10 year lag between stratification and the collection of the survey data, which would be roughly the average lag time for the two-phase CPS. The variables used were number of unemployed persons and number of persons in the civilian labor force. The ratios were computed for 31 states. Averages of these ratios over these 31 states were also computed. The remaining states were omitted for various reasons, as described in Ernst (1990).

When computing the variances, the number of sample persons was first obtained for the independent sample method to meet the proposed D_2 reliability requirements. For each of the other three procedures, the same number of sample persons was assumed. For each of these four methods, the within PSUs variances were obtained by computing the simple random sampling with replacement variance for that size sample and multiplying by a design factor to account for the fact that clustered, systematic sampling was actually used within each PSU. For the multiple WLs method, this approach to computing the within PSUs variances is at least partially justified by the results at the end of the previous section. The within PSUs component of each variance

is thus computed to be the same for all four methods and the differences among the methods are due solely to differences in the between PSUs component for all methods, and also the between strata component for the controlled selection and multiple WLs methods, which are the only methods among the four methods to have such a component.

From Table 1 it can be observed that the variances for the multiple WLs method are generally less than those for the independent supplement method, but higher than those for independent selection and controlled selection. These results are not surprising. In both controlled selection and independent selection, the PSUs are selected from an optimal D_2 stratification, and therefore these methods would be expected to result in lower variances than multiple WLs, which selects all its PSUs from a stratification that is optimal for D_1, not D_2.

Lower variances for multiple WLs than for independent supplement can be attributed to the fact that multiple WLs constrains the actual number of WLs selected from each D_1 stratum to be within one of the expected number, while independent supplement does not. As a result, comparisons between variances for these two methods should be analogous to comparisons between variances for without replacement and with replacement sampling.

Although independent selection and controlled selection result in lower variances than multiple WLs, each of these methods has a major drawback. Independent selection generally does not retain all D_1 sample PSUs in the D_2 sample. Controlled selection requires that the D_1 and D_2 PSUs be selected simultaneously, and therefore cannot be used for an expansion planned after the D_1 sample is in place. Consequently, multiple WLs and independent supplement may be the only methods among these four that are operationally feasible.

Because the within PSUs component of variance generally is the dominant component of variance for CPS, the ratios in Table 1 usually differ little from 1. In Table 2, the same ratios are presented, with the within PSUs component omitted from each variance. The ordering of the relationships, of course, remains unchanged, but there are larger deviations from 1.

6. References

Chandhok, P., Weinstein, R., and Gunlicks, C. (1990). Augmenting a Sample to Satisfy Subpopulation Reliability Requirements. Proceedings of the Section on Survey Research Methods, American Statistical Association, 696-701.

Cochran, W.G. (1977). Sampling Techniques. New York: John Wiley and Sons.

Ernst, L.R. (1990). Simultaneous Selection of Primary Sampling Units for Two Designs. Proceedings of the Section on Survey Research Methods, American Statistical Association, 688-693.

Sampford, M.R. (1967). On Sampling Without Replacement With Unequal Probabilities of Selection. Biometrika, 54, 499-513.

Table 1

Ratios of Total Variances for Other Options
to the Independent Sample

State	Unemployed			Civilian Labor Force		
	CS	IS	MW	CS	IS	MW
Alabama	1.001	1.019	1.016	1.000	1.028	1.064
Arizona	1.000	1.003	1.014	0.998	1.069	1.016
Arkansas	1.001	1.032	0.991	0.999	1.036	0.957
Colorado	1.000	1.031	1.024	1.000	1.212	1.065
Georgia	1.000	1.010	1.002	1.000	1.024	0.990
Idaho	1.002	1.259	1.041	1.003	1.155	1.045
Indiana	0.999	1.028	1.033	1.018	1.112	1.003
Iowa	0.998	1.030	1.001	0.997	1.055	1.001
Kansas	1.000	1.026	1.009	0.996	1.035	0.993
Kentucky	1.001	1.020	1.028	0.997	1.050	1.040
Louisiana	1.000	1.007	1.012	0.999	1.054	1.038
Maryland	1.000	1.020	1.000	1.000	1.196	0.952
Minnesota	1.002	1.011	1.014	0.996	1.057	1.013
Mississippi	0.999	1.047	1.003	1.001	1.132	0.980
Missouri	1.000	1.026	1.010	1.005	1.120	0.979
Montana	0.997	1.032	1.043	1.009	1.261	0.979
Nebraska	1.000	1.020	1.006	1.000	1.078	0.983
Nevada	1.000	1.035	0.998	0.992	1.855	0.960
New Mexico	0.999	1.013	1.040	1.011	1.084	1.053
North Dakota	0.999	1.071	1.011	0.986	1.106	0.978
Oklahoma	1.000	1.012	1.009	0.993	1.117	0.995
Oregon	0.999	1.015	1.021	0.999	1.216	0.981
South Carolina	1.003	1.054	0.993	1.002	1.135	1.009
South Dakota	0.996	1.054	1.020	0.997	1.092	0.977
Tennessee	1.001	1.047	1.005	1.000	1.060	0.993
Utah	1.000	1.019	1.002	0.995	1.128	0.980
Virginia	0.999	1.031	0.998	1.018	1.112	0.970
Washington	0.999	1.034	1.012	1.008	1.148	1.015
West Virginia	1.000	1.002	1.012	1.001	0.994	1.002
Wisconsin	0.999	1.028	1.015	0.999	1.086	0.989
Wyoming	0.999	1.014	1.012	1.000	1.212	1.067
Mean	1.000	1.034	1.013	1.001	1.130	1.002

CS = Controlled Selection
IS = Independent Supplement
MW = Multiple Workloads

Table 2

Ratios of Between PSU Variances for Other Options to the Independent Sample

State	Unemployment			Civilian Labor Force		
	CS	IS	MW	CS	IS	MW
Alabama	1.12	3.36	2.97	1.03	2.87	5.33
Arizona	1.07	4.21	18.63	0.89	4.68	1.84
Arkansas	1.06	2.67	0.53	0.99	1.57	0.33
Colorado	1.01	4.32	3.59	1.00	5.31	2.33
Georgia	1.04	2.42	1.25	1.02	1.91	0.64
Idaho	1.10	13.47	2.96	1.06	4.64	2.05
Indiana	0.89	4.00	4.52	1.47	3.90	1.08
Iowa	0.78	3.63	1.10	0.83	4.29	1.09
Kansas	0.95	3.49	1.89	0.92	1.71	0.85
Kentucky	1.08	3.35	4.21	0.84	3.67	3.13
Louisiana	0.95	2.65	3.88	0.97	2.63	2.10
Maryland	1.00	4.62	1.09	1.00	4.92	0.04
Minnesota	1.12	1.51	1.69	0.87	2.75	1.39
Mississippi	0.93	4.84	1.28	1.03	3.70	0.59
Missouri	0.95	3.99	2.17	1.09	3.25	0.61
Montana	0.88	2.16	2.53	1.19	6.24	0.58
Nebraska	0.99	2.83	1.52	1.00	2.78	0.62
Nevada	1.02	5.81	0.78	0.86	15.02	0.35
New Mexico	0.91	3.06	7.27	1.41	4.13	2.98
North Dakota	0.91	6.34	1.85	0.72	3.09	0.57
Oklahoma	1.02	2.72	2.27	0.83	3.86	0.88
Oregon	0.89	2.25	2.72	0.98	5.40	0.61
South Carolina	1.17	3.64	0.68	1.10	9.67	1.56
South Dakota	0.90	2.53	1.56	0.95	2.50	0.63
Tennessee	1.10	8.16	1.77	1.00	3.01	0.76
Utah	0.97	2.66	1.19	0.94	2.67	0.73
Virginia	0.95	2.87	0.87	1.33	3.07	0.44
Washington	0.94	3.29	1.78	1.25	5.86	1.48
West Virginia	0.93	1.60	4.23	1.12	0.37	1.22
Wisconsin	0.96	2.17	1.62	0.98	2.94	0.75
Wyoming	0.67	5.69	5.04	1.03	35.91	12.08
Mean	0.98	3.88	2.89	1.02	5.11	1.60

CS = Controlled Selection
IS = Independent Supplement
MW = Multiple Workloads