MULTIPLE WORKLOADS PER STRATUM SAMPLING DESIGNS

Lynn Weidman and Lawrence R. Ernst*
Bureau of the Census, Washington, DC 20233

1. INTRODUCTION

The Current Population Survey (CPS) is a multi-stage sampling design with primary sampling units (PSUs) selected with probability proportional to size and ultimate units sampled so that each has the same overall probability of selection. The PSUs are stratified so that the number of ultimate units sampled will be approximately the same in each stratum. PSUs that are too large to fit into these strata are selected with probability one (certainty PSUs). In a one-PSU-per-stratum (1PS) design like CPS, the probability of a noncertainty PSU being in sample is p_{ij}=Size(PSU_{ij})/Size(stratum i), where PSU_{ij} is the j^{th} PSU in stratum i. The number of ultimate units chosen in stratum i is the same regardless of which PSU is chosen, and this is the workload (WL) of a single field interviewer. We can translate the probability of PSU selection into the expected number of WLs selected to be in a PSU.

$$E(WLs \text{ in } PSU_{ij}) = 0(1-p_{ij}) + (1)p_{ij} = p_{ij}, \qquad (1)$$

which is the probability that PSU_{ij} is selected. This shows that instead of considering selection probabilities we can think in terms of WLs in 1PS designs.

Next extend this expected WL approach to two-PSUs-per-stratum (2PS) designs. When sampling without replacement with any of the widely used procedures, the probability of noncertainty PSU_{ij} being selected is $2p_{ij}$. These procedures are applicable only if $2p_{ij}<1$ for all PSUs. However, as we will show, even if this inequality is not satisfied we can select PSUs to receive each of the two WLs in such a way that E(WLs in PSU_{ij}) = $2p_{ij}$ for all PSUs. Furthermore, for our method, if $2p_{ij}<1$, PSU_{ij} gets either 0 or 1 WLs; if $2p_{ij}>1$, PSU_{ij} gets 1 or 2 WLs; and if $2p_{ij}=1$, PSU_{ij} always gets 1 WL. This approach minimizes between PSUs variance, and we call these 2-WLs-per-stratum (2WL) designs.

In general this idea can be extended to an s-WLs-per-stratum (sWL) design, in which the number of WLs that a noncertainty PSU can receive varies by at most 1. Let$\lfloor x \rfloor$ represent the largest integer not exceeding x and$\lceil x \rceil = \lfloor x \rfloor + 1$. Then the general WL restrictions for an sWL design are:

S.1 If sp_{ij} is an integer, PSU_{ij} always gets sp_{ij} WLs.

S.2 If sp_{ij} is not an integer, PSU_{ij} can get either $\lfloor sp_{ij} \rfloor$ or $\lceil sp_{ij} \rceil$ WLs.

The probabilities of PSU_{ij} getting $\lfloor sp_{ij} \rfloor$ and $\lceil sp_{ij} \rceil$ WLs in S.2 are determined so that $E(WLs$ in $PSU_{ij}) = sp_{ij}$, for all ij.

We have derived procedures for selecting PSUs to receive WLs in 2WL and 3WL designs, since those are the situations that occur in the present application. It should be possible to develop procedures for more than 3 WLs per stratum using similar approaches, although with increasing complexity as the number of WLs increases.

This work was motivated by a formerly planned expansion of the CPS that would 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. See Ernst (1990) and Chandhok, Weinstein and Gunlicks (1990) for other options. Each of the other approaches 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 a double workload 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 only involves selecting PSUs to receive the additional sample. In order to accomplish this, the concept of multiple-PSUs-per-stratum designs was generalized to multiple-workloads-per-stratum designs. Although the CPS application motivated this work, there are potential applications to other sample expansion problems.

Due to space limitations, some mathematical details, derivations and two appendices have been excluded from this version of our paper. The complete paper may be obtained by writing to the authors (Weidman and Ernst (1991)).

2. EXPANDING AN EXISTING DESIGN

In this application we denote the current CPS design by D_1 and an expanded CPS with additional sample by D_2. We have been looking into options for stratifying and selecting noncertainty PSUs in D_1 that maintain as many PSUs in sample as possible when expanding D_1 to D_2. One way of ensuring that $D_1 \subset D_2$ is to begin with D_1 and then select additional WLs from the D_1 strata. These WLs are selected to meet the general requirements for multiple-WLs-per-stratum designs given in section 1 . The sampling intervals for D_1 and D_2 are SI_1 and SI_2. SI_1 is determined by finding a stratification of PSUs for a 1PS design that attains specified variance requirements while keeping stratum WL sizes within acceptable bounds. SI_2 is similarly calculated.

When expanding from D_1 to D_2, the number of additional WLs needed per stratum must be calculated. The total sample size for D_2 is $R=SI_1/SI_2$ times the sample size for D_1, so the total number of WLs required is also R times the number of WLs in D_1. Since there is one WL per stratum in D_1, the expected number of WLs per stratum in D_2 will be R. If R is not integral, the between strata portion of between PSUs variance will be minimized by letting each stratum have either $\lfloor R \rfloor$ or $\lceil R \rceil$ WLs. Giving each stratum $\lceil R \rceil$ WLs with probability $R-\lfloor R \rfloor$ and $\lfloor R \rfloor$ WLs with probability $\lceil R \rceil-R$ gives R as the expected number of WLs per stratum.

Let I denote the number of strata in the designs. Then ideally $I(R-\lfloor R \rfloor)$ strata get $\lceil R \rceil$ WLs and $I(\lceil R \rceil-R)$ strata get $\lfloor R \rfloor$ WLs in D_2. Of course $I(R-\lfloor R \rfloor)$ will not usually be an integer, so generally we will round up the number of strata with $\lceil R \rceil$ WLs and round down the number with $\lfloor R \rfloor$ WLs to ensure that the required sample size is attained. In the case that R itself is an integer or slightly less than an integer, we would use that integral number of WLs for each stratum. Let $T=\lfloor IR \rfloor+1\approx IR$ be the total number of WLs in D_2 and denote by $Q_{\lfloor R \rfloor}=\lceil R \rceil-T/I$ the actual probability of a stratum getting $\lfloor R \rfloor$ WLs. Letting w_{ij} be the number of WLs assigned to PSU_{ij}, then $E(w_{ij})$ is

$$Q_{\lfloor R \rfloor}\lfloor R \rfloor p_{ij} + (1-Q_{\lfloor R \rfloor})(\lfloor R \rfloor+1)p_{ij} = p_{ij}T/I.$$

If SI_{ij} is the sampling interval used to obtain a single WL within this PSU, its value is obtained from the relationship

$$p_{ij}T/[I(SI_{ij})] = 1/SI_2 = R/SI_1. \qquad (2)$$

Solving for SI_{ij} gives $SI_{ij} = p_{ij}(SI_1)T/IR \approx p_{ij}SI_1$.

Since $p_{ij}SI_1$ is the within PSU sampling interval for D_1, using this approach gives the desired result that the workload sizes in a stratum for the two designs are about equal. Those for D_2 are slightly smaller when R is not integral, since then T/I is slightly larger than R.

(In practice the determination of SI_1 and SI_2 will not be as simple as indicated here. For the CPS application an approximate SI_1 is calculated to be used as a parameter in an initial stratification of the PSUs in a state. The actual variances of variables of interest are computed and SI_1 is adjusted to get a minimum sample size that more closely meets the variance requirements. This procedure is repeated until a satisfactory stratification is arrived at. This iterative approach is necessary because of the relative contributions of the between and within PSUs components, which vary with each stratification and the ratio of the within PSU variance to the total variance. For D_2 we first calculate the variances of interest using an approximate SI_2 and the final stratification for D_1. If the variance requirements are not met, SI_2 is decreased by a proportion which again depends upon the relative sizes of the variance components. If the variances are smaller than required, SI_2 is increased to get a smaller sample. This procedure is iterated until a minimum sample size that meets the variance requirements for D_2 is determined.)

3. SELECTING PSUs TO RECEIVE WORKLOADS

We are now left with devising a sampling procedure which will give us a one-PSU(WL)-per-stratum D_1 whose selected PSUs are included in a multiple-WL-per-stratum D_2. As mentioned previously, we have derived procedures for 1, 2 and 3 WLs per stratum, since they are the situations we encounter in our application. A starting point for procedures of this type are methods of Brewer-Durbin (Cochran, 1977) for 2PS designs and Sampford (1967) for 3+PS designs, which we use when applicable. For simplicity of notation consider a single stratum with PSUs having probability of selection $p_1 \geq p_2 \geq p_3 \ldots \geq p_J$ in D_1, where p_j=Size(PSU_j)/Size(stratum). We will look at all the selection possibilities for D_2, using the notation

$P(i|j)$ = $P(PSU_i$ gets 2nd $WL|PSU_j$ got 1st WL),
$P(jk|i)$ = $P(PSU_j$ and PSU_k get 2nd and 3rd WLs in any order$|PSU_i$ got 1st WL),
$P(ij)$ = $P(PSU_i$ and PSU_j each get 1 WL in any order),
$P(ijk)$ = $P(PSU_i$, PSU_j and PSU_k each get 1 WL in any order).

The following are also used in devising procedures for sWL designs.

(a) In deriving the conditional probabilities it will be assumed that each possible order of selection for a set of PSUs is equally likely.

(b) $P(PSU_i \text{ gets } \lceil sp_i \rceil\ WLs)=sp_i-\lfloor sp_i \rfloor$

(c) $P(PSU_i \text{ gets } \lfloor sp_i \rfloor\ WLs)=\lceil sp_i \rceil-sp_i$.

It is easy to see that with these probabilities $E(WLs$ in $PSU_i) = sp_i$.

A. Select 2 WLs for a stratum

1. $p_1 \geq 1/2$
   $P(1|1) = 2-p_1^{-1}$
   $P(k|1) = p_k/p_1,\ k \neq 1$
   $P(1|k) = 1$

   Derivation:
   $P(11) = 2p_1-1$ by (b) and hence $P(1|1) =$
   $(2p_1-1)/p_1 = 2-p_1^{-1}$
   $P(1|k) = 1$ since PSU_1 always gets at least 1 WL
   Since $P(PSU_k$ gets 2nd WL) = p_k by (a) and
   $P(k|\bar{1})=0,\ P(k|1) = p_k/p_1$

(Further derivations are omitted due to lack of space.)

2. $p_1<1/2$
   Use Durbin procedure

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

B. Select 3 *WLs* for a stratum

1. $p_1 \geq 2/3$
   $P(11|1) = (3p_1-2)/p_1$
   $P(1k|1) = 2p_k/p_1,\ k \neq 1$
   $P(11|k) = 1$

2. $p_1, p_2 > 1/3$
   $P(11|2) = (3p_1-1)/3p_2 \quad P(22|1) = (3p_2-1)/3p_1$

   $P(12|2) = 2(3p_2-1)/3p_2 \quad P(12|1) = 2(3p_1-1)/3p_1$
   $P(1k|2) = p_k/p_2\ k \neq 1,2 \quad P(2k|1) = p_k/p_1$
   $P(12|k) = 1$

3. $p_1 < 1/3$
   Use $P(jk|i) = K_3 p_j p_k \times$

$$\left[\frac{1}{(1-3p_i)(1-3p_j)}+\frac{1}{(1-3p_i)(1-3p_k)}+\frac{1}{(1-3p_j)(1-3p_k)}\right]$$

See Sampford (1967) for his procedure and definition of K_3.

4. $1/3 \leq p_1 < 2/3,\ p_2 < 1/3$

   a. $p_2/(1-p_1) \geq 1/2$

$$P(12|1)=\frac{2(3p_1+3p_2-2)}{3p_1} \qquad P(1k|2)=\frac{(2-3p_1)p_k}{3p_2(1-p_1-p_2)}$$

$$P(1k|1)=\frac{2(1-3p_2)p_k}{3p_1(1-p_1-p_2)} \qquad P(11|k)=\frac{(1-3p_2)p_k}{3p_k(1-p_1-p_2)}$$

$$P(2k|1)=\frac{(2-3p_1)p_k}{3p_1(1-p_1-p_2)} \qquad P(12|k)=\frac{(2-3p_1)}{(1-p_1-p_2)}$$

$$P(11|2)=\frac{3p_1+3p_2-2}{3p_2}$$

   b. $p_2/(1-p_1) < 1/2$

$$P(1k|1)=\frac{2(3p_1-1)p_k}{3p_1(1-p_1)} \qquad P(11|k)=\frac{(3p_1-1)}{3(1-p_1)}$$

$$P(jk|1)=\frac{(2-3p_1)D(j,k)}{3p_1} \qquad P(1j|k)=\frac{(2-3p_1)D(j,k)}{3p_k}$$

$$D(j,k)=2p_j'p_k'\left[\frac{1}{1-2p_j'}+\frac{1}{1-2p_k'}\right] \Big/ \left[1+\sum_{i=2}^{J}\frac{p_i'}{1-2p_1'}\right]$$

and $p_j'=p_j/(1-p_1)$.

4. VARIANCE DECOMPOSITION

Let $S = (s_1, s_2, \ldots, s_I)$ represent a combination of number of WLs per stratum in a design with I strata and T WLs. There are three stages of sampling: (1) selection of S; (2) selection of PSUs to receive WLs within each stratum; and (3) selection of ultimate units within PSUs. We want to calculate the component of variance due to each of these stages.

Since the within PSU component of variance typically has the largest contribution to the variance in household surveys conducted by the Census Bureau, and equal weighting of all ultimate sampling units minimizes this component under common simplifying assumptions, our sample estimator of population total, $\hat{y}$, for D_2 is obtained by simply summing all D_2 sample units and multiplying by SI_2. This is clearly an unbiased estimator.

We use here the notation of section 2. For any PSU_{ij} with $w_{ij}>0$ in a given sample, let $\hat{y}_{ij}$ be the estimator obtained by summing the value of y for all sample units in PSU_{ij} and multiplying by SI_{ij}/w_{ij}. Given $w_{ij}>0$, $\hat{y}_{ij}$ is an unbiased estimator of the total for PSU_{ij}, y_{ij}. If $w_{ij} = 0$, let $\hat{y}_{ij}=0$. An estimate of the population total from a chosen sample t is then

$$\hat{y}_t=\sum_{i=1}^{I}\sum_{j=1}^{N_i}\frac{w_{ij}SI_2}{SI_{ij}}\hat{y}_{ij}, \tag{3}$$

where all values of w_{ij} and $\hat{y}_{ij}$ depend upon t. From (2), $SI_{ij} = p_{ij}T(SI_2)/I$, so we write the general estimate of total y_t as

$$\hat{y}_t=\sum_{i=1}^{I}\sum_{j=1}^{N_i}a_{ij}\hat{y}_{ij}, \tag{4}$$

where

$$a_{ij}=Iw_{ij}/(Tp_{ij}) \tag{5}$$

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

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

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

$$+E_1E_2V_3\left(\sum_{i=1}^{I}\sum_{j=1}^{N_i}a_{ij}\hat{y}_{ij}\right) \tag{6}$$

The subscripts on the expectations denote the 3 stages of sampling. If within PSU sampling is carried out so that whenever $w_{ij}>0$, $E_3(\hat{y}_{ij})=y_{ij}$, the PSU_{ij} total, then

$$V(\hat{y}) = V_1\left[\sum_{i=1}^{I}\sum_{j=1}^{N_i} E_2(a_{ij})y_{ij}\right] + E_1\left[\sum_{i=1}^{I} V_2\left(\sum_{j=1}^{N_i} a_{ij}y_{ij}\right)\right]$$

$$+E_1\left[\sum_{i=1}^{I}\sum_{j=1}^{N_i} E_2\{a_{ij}^2V_3(\hat{y}_{ij})\}\right] \qquad (7)$$

These three terms are, respectively, the between strata, between PSUs within strata and within PSUs components of variance. Note that a_{ij} depends on $s_1,s_2,...,s_I$ and w_{ij}.

4.1 Some Moments of the a_{ij}

In order to expand the expressions for the variance components, the expectations $E_2(a_{ij})$, $E_2(a_{ij}^2)$ and $E_2(a_{ij}a_{ik})$ for the second sampling stage are needed. The number of WLs in stratum i is s_i, which is determined from the first sampling stage. For the second stage we treat stratum i as a s_iWL design, so for all ij, $E_2(w_{ij}) = s_ip_{ij}$.

$$w_{ij} = \begin{cases} \lfloor s_ip_{ij}\rfloor \textit{ with probability } \lceil s_ip_{ij}\rceil - s_ip_{ij} \\ \lceil s_ip_{ij}\rceil \textit{ with probability } s_ip_{ij} - \lfloor s_ip_{ij}\rfloor \end{cases}$$

$$E_2(a_{ij}) = s_iI/T \qquad (8a)$$

$$E_2(a_{ij}^2) = \frac{I^2}{T^2p_{ij}^2}\left[(\lfloor s_ip_{ij}\rfloor)(2s_ip_{ij} - \lfloor s_ip_{ij}\rfloor - 1) + s_ip_{ij}\right] \qquad (8b)$$

$$E_2(a_{ij}a_{ik}) = \frac{I^2}{T^2p_{ij}p_{ik}}\sum_{w_{ij}\varepsilon U_{s_i}}\sum_{w_{ik}\varepsilon V_{s_i}} w_{ij}w_{ik}Q_{ijks_i}(w_{ij},w_{ik}) \qquad (8c)$$

where U_{s_i} is the set of possible numbers of WLs that PSU_{ij} can have when the stratum gets s_i WLs, V_{s_i} is the corresponding set for PSU_{ik}, and $Q_{ijks_i}(w_{ij},w_{ik})$ is the probability that PSU_{ij} and PSU_{ik} simultaneously get w_{ij} and w_{ik} WLs, given s_i. (See Appendix 1 of the complete paper for the joint selection probabilities of WLs for the various sampling procedures.) If s_ip_{ij} is an integer, these last two expectations each simplify to $s_i^2I^2/T^2$.

4.2 Between Strata Variance

If each stratum gets the same number of WLs, this component is zero. This section looks at the general case where each stratum is assigned either $\lfloor R\rfloor$ or $\lfloor R\rfloor+1$ WLs in the first sampling stage.

Recall that $V_1\left[\sum_{i=1}^{I}\sum_{j=1}^{N_i} E_2(a_{ij})y_{ij}\right]$ is the between strata component of variance, and from (8a) the summation is

$$\sum_{i=1}^{I}\sum_{j=1}^{N_i}\frac{s_iI}{T}y_{ij} = \frac{I}{T}\sum_{i=1}^{I} s_iy_{i.}\,.$$

In order to simplify the notation that follows, let $L = \lfloor R\rfloor$ and X = number of strata that get $L+1$ WLs, so that each stratum has probability X/I of getting $L+1$ WLs in the first sampling stage and probability $(I\text{-}X)/I$ of getting L WLs. The total number of WLs in the design is then $T = IL+X$. Now

$$E_1\left(\frac{I}{T}\sum_{i=1}^{I} s_iy_{ij}\right) = \frac{I}{T}\sum_{i=1}^{I} E_1(s_i)y_{i.} = \frac{I}{T}\sum_{i=1}^{I}\frac{T}{I}y_{i.} = \sum_{i=1}^{I} y_{i.}\,.$$

Then

$$V_1\left(\frac{I}{T}\sum_{i=1}^{I} s_iy_{ij}\right) = E_1\left[\sum_{i=1}^{I}\frac{I}{T}s_iy_{i.} - \sum_{i=1}^{I} y_{i.}\right]^2$$

$$= E_1\left[\sum_{i\varepsilon S_{L+1}}\frac{I}{T}y_{i.} - \sum_{i=1}^{I}\left(1-\frac{IL}{T}\right)y_{i.}\right]^2$$

where S_{L+1} is the set of strata where $s_i = L+1$.

Using $X=T\text{-}IL$ we can rewrite this as

$$\frac{I^2X^2}{T^2}E_1\left[\sum_{i\varepsilon S_{L+1}}\frac{y_{i.}}{X} - \sum_{i=1}^{I}\frac{y_{i.}}{I}\right]^2. \qquad (9)$$

Since in any sample S, the set S_{L+1} is a simple random sample of X out of I strata, this is $(IX/T)^2$ times the variance of the mean from a simple random sample. Using Theorem 2.2 of Cochran (1977), we obtain that (9) reduces to

$$\frac{I^2X^2}{T^2}\,\frac{\sum_{i=1}^{I}(y_{i.}-\bar{y}_{..})^2}{X(I-1)}\left(\frac{I-X}{I}\right) = \frac{(I-X)IX}{T^2(I-1)}\left[\sum_{i=1}^{I} y_{i.}^2 - \frac{y_{..}^2}{I}\right] \qquad (10)$$

4.3 Between PSUs Within Strata Variance

The term $E_1\left[\sum_{i=1}^{I} V_2\left(\sum_{j=1}^{N_i} a_{ij}y_{ij}\right)\right]$ in (7) represents the between PSUs within strata component of variance and can be expanded as

$$E_1\left\{\sum_{i=1}^{I}\left[\sum_{j=1}^{N_i} E_2(a_{ij}^2)y_{ij}^2 + \sum_{j=1}^{N_i}\sum_{\substack{k=1\\ j\neq k}}^{N_i} E_2(a_{ij}a_{ik})y_{ij}y_{ik}\right]\right\}$$

$$-E_1\left\{\sum_{i=1}^{I}\left[\sum_{j=1}^{N_i} E_2(a_{ij})y_{ij}\right]^2\right\}. \qquad (11)$$

The first term cannot generally be simplified, but the appropriate $E_2(a_{ij}^2)$ and $E_2(a_{ij}a_{ik})$ are given in (8b) and (8c). The final term is

$$E_1\left\{\sum_{i=1}^{I}\left(\frac{s_iI}{T}y_{i.}\right)^2\right\} = \frac{I^2}{T^2}\sum_{i=1}^{I} E_1(s_i^2)y_{i.}^2 = \frac{I(IL^2+2XL+X)}{T^2}\sum_{i=1}^{N_i} y_{i.}^2. \tag{12}$$

4.4. Within PSUs Variance

Now we will look at the variability of estimates of PSU totals due to within PSU sampling assuming all sample WLs within a PSU are selected by simple random sampling without replacement. Appropriate modifications are necessary for other within PSU selection procedures. Recall that the within PSUs variance component is $E_1\left[\sum_{i=1}^{I}\sum_{j=1}^{N_i} E_2\{(a_{ij}^2)V_3(\hat{y}_{ij})\}\right]$.

The contribution of PSU_{ij} to this expectation can be written as

$$\sum_{z\epsilon S} P(s_i{=}z) \sum_{w=\lfloor s_ip_{ij}\rfloor}^{\lceil s_ip_{ij}\rceil} P(w_{ij}{=}w\,|\,s_i{=}z)\times \left(\frac{I}{Tp_{ij}}\right)^2 \frac{wM_{ij}(M_{ij}-wm_{ij})S_{3ij}^2}{m_{ij}} \tag{13}$$

where

M_{ij} = size of PSU_{ij} ,
$m_{ij} = m_i$ = WL size in stratum i,
S_{3ij}^2 = variance of y in PSU_{ij} ,
$\dfrac{M_{ij}(M_{ij}-wm_{ij})}{wm_{ij}}S_{3ij}^2$ is the variance of an estimate of y_{ij} from a simple random sample without replacement of size wm_{ij}.

If the finite population factor is negligible in (13), then we have that (13) is approximately

$$\sum_{z\epsilon S} P(s_i{=}z) \sum_{w=\lfloor s_ip_{ij}\rfloor}^{\lceil s_ip_{ij}\rceil} P(w_{ij}{=}w\,|\,s_i{=}z)w\left(\frac{IM_{ij}}{Tp_{ij}}\right)^2\frac{S_{3ij}^2}{m_{ij}}. \tag{14}$$

Since

$$\sum_{z\epsilon S} P(s_i{=}z) \sum_{w=\lfloor s_ip_{ij}\rfloor}^{\lceil s_ip_{ij}\rceil} P(w_{ij} = w\,|\,s_i = z)w =$$

$$E_1[E_2(w_{ij}|s_i)] = E_1(p_{ij}s_i) = p_{ij}T/I,$$

and from (2)

$$IM_{ij}/(Tp_{ij}m_i) = I(SI_{ij})/(Tp_{ij}) = SI_2,$$

(14) simplifies to

$$IM_{ij}^2S_{3ij}^2/(Tp_{ij}m_i) = (SI_2)M_{ij}S_{3ij}^2. \tag{15}$$

Finally, if the S_{3ij}^2 are the same for all ij, with common value denoted S_3^2, then by summing (15) over all ij we obtain that the within PSU variance is approximately $(SI_2)MS_3^2$, where $M=\sum_{i=1}^{I}\sum_{j=1}^{N_i} M_{ij}$. This is the sampling variance for the standard estimator of population total from a simple random sample with replacement, for a sample of size $M/(SI_2)$ selected from a population of size M with variance S_3^2. Similar assumptions lead to the same approximate within PSU variance for the other options investigated for two phase sampling (mentioned in section 1), 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 workloads 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 D_2 independently from a second, supplemental stratification.

In Table 1 the ratio of variances for controlled selection, independent supplement, and multiple workload methods, to the independent supplement method are presented. For all four methods 1980 census data were used to obtain the stratification, 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).

From Table 1 it can be observed that the variances for the multiple workloads method are generally less than those for the independent supplement method, but considerably 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 workloads, which selects all its PSUs from a stratification that is optimal for D_1, not D_2.

Lower variances for multiple workloads than independent supplement can be attributed to the fact that multiple workloads constrain the actual number of PSUs 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 workloads, 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 workloads and independent supplement may be the only methods among these four that are operationally feasible.

In 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 PSU 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 workloads method this approach to computing the within PSU variances is at least partially justified by the results at the end of the previous section.

The within PSU component of each variance is thus computed to be the same for all four methods and the differences among total variances for the methods are due solely to differences in the between PSU component and the between strata component for the multiple workloads method, which is the only one of the four methods to have such a component. For a given survey the effect of any of these options on total variance depends upon the relative magnitudes of the within and between PSUs variances.

Because the within PSUs component of variance generally is the dominant component of variance for CPS, the differences in between PSUs variance shown in Table 1 have little effect on total variance. Table 2 shows this for the means of the states. The state values show as little variance between methods as do these means.

REFERENCES

Chandhok, P., Weinstein, R., and Gunlicks, C. (1990), "Augmenting a Sample to Satisfy Subpopulation Reliability Requirements," in *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 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," in *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 688-693.

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

Weidman, L. and Ernst, L.R. (1991), "Multiple Workloads Per Stratum Sampling Designs," Report CENSUS/SRD/RR-91/10, U.S. Bureau of the Census, Statistical Research Division.

Table 1
Ratios of Between PSU Variances for Other
Options to the Independent Supplement

	Unemployed			Civilian Labor Force		
State	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
N 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
S Carolina	1.17	3.64	0.68	1.10	9.67	1.56
S 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
W. 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

Table 2
Ratios of Total Variances for
Other Options to the Independent Supplement

	Unemployed			Civilian Labor Force		
	CS	IS	MW	CS	IS	MW
Mean	1.00	1.03	1.01	1.00	1.13	1.00

* This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author(s) and do not necessarily reflect those of the Census Bureau.