VARIANCE ESTIMATION FOR THE NATIONAL COMPENSATION SURVEY WHEN PSUS ARE CLUSTERED PRIOR TO THE SECOND PHASE OF SAMPLING

Christopher J. Guciardo, Alan H. Dorfman, Lawrence R. Ernst, Michail Sverchkov
Guciardo.Chris@bls.gov, Dorfman.Alan@bls.gov, Ernst.Lawrence@bls.gov, Sverchkov.Michael@bls.gov
Bureau of Labor Statistics, 2 Massachusetts Ave., N.E., Room 3160, Washington, DC 20212-0001

KEY WORDS: Replication; zero-calibrated; BRR; jackknife; NCS

1. Introduction

In this paper we investigate the accuracy of the current variance estimator used in the National Compensation Survey (NCS) and some design-based alternatives. The principal motivation for this study is that our current variance estimator is based in part on the assumption, which is not completely true in NCS, that the sampling of establishments is done independently in each sample area.

Three of the Bureau of Labor Statistics (BLS) compensation survey programs, the Employment Cost Index (ECI), the Employee Benefits Survey (EBS), and the locality wage surveys, were integrated, creating one comprehensive National Compensation Survey (NCS) program. The ECI publishes national indexes that track quarterly and annual changes in employers' labor costs, both wages and benefits, and also cost level information, previously annually but now quarterly, on the cost per hour worked of each component of compensation. The cost level counterpart of the ECI is referred to as Employee Costs for Employee Compensation (ECEC). The former EBS publishes annual incidence and detailed provisions of selected employee benefit plans and is now known as the Benefits Incidence and Provisions Product. The locality wage surveys program publishes locality and national occupational wage data. There are two samples used in the NCS program, a larger wage sample, used for those survey products that only present wage data, and a smaller wage and benefits subsample used for products that include benefits data. We will only be considering the wage sample in this paper.

The NCS wage sample mostly consists of five rotating panels at any point in time. It is selected using a three-stage stratified design with probability proportionate to employment sampling at each stage. The first stage of sample selection is a sample of 154 areas (PSUs); the second stage is a sample of establishments within sampled areas, with the frame of establishments presently partitioned into 23 industry sampling strata; and the third stage is a sample of occupations within sampled areas and establishments. Currently, government and aerospace manufacturing establishments, although included in the survey estimates, are not part of the rotating panel design.

A key aspect of the sample design is that the second stage of sampling, the selection of establishments, is not done independently in each of the 154 sample areas. If we had sampled independently in each sample area × industry sampling stratum cell, then the total number of sampling cells would be 154×23=3542 (since there are 23 industry sampling strata). Since we require a minimum of one unit in each nonempty cell, the total number of sample units to meet the minimums would then comprise a large proportion of each rotating sample panel, resulting in an inefficient design. To alleviate this problem, we only sample independently in 52 larger areas. The three non-metropolitan areas from Alaska and Hawaii are collapsed to form one cluster; and the remaining 99 areas are collapsed into a second, much larger cluster. The reason for the two separate clusters is that the NCS wage estimates are used in determining locality pay for federal workers, but only in the 48 contiguous states and the District of Columbia, which led us to sample Alaska and Hawaii separately from the remainder of the nation. For each of these clusters, there is a single allocation for each industry stratum to all the PSUs in the cluster. The set of frame establishments in the cluster is sorted first by PSU and then by the frame employment. A systematic PPS sample of establishments is then selected, with the measure of the size being the product of the frame employment and the reciprocal of the probability of selection of the PSU in which the establishment is located.

Our current method of calculating variance estimates uses the variation of balanced repeated replication (BRR) developed by Fay (1989). BRR assumes that the sample design is two PSUs per stratum with replacement. Since our design is actually one PSU per stratum, we artificially created a two PSU per stratum design by collapsing, for the most part, pairs of noncertainty PSUs together to form what we call variance strata. This is a common technique know as collapsed stratum variance estimation. The certainty PSUs were split in half to form additional variance strata, with half of the noncertainty sample establishments in each industry sampling stratum in one variance PSU and half in the other, and with the certainty establishments in both variance PSUs. Because the variance strata that we formed do not completely reflect our actual design, it should be anticipated that our variance estimator is biased. The bias arising from the use of a collapsed stratum variance estimator has been well studied (Wolter 1985). In addition, it is assumed in BRR that sampling is done independently in each variance PSU. This is not the case for the two multi-area clusters and as a result there is another source of bias, which will be the major focus of the paper.

This issue was studied previously in Wang, Dorfman, and Ernst (2004), where a hybrid variance estimator dependent on both model-based and designed-based ideas was introduced as an alternative to our current variance estimator. Although that paper developed from the NCS problem, the empirical part of the paper was based on an

artificial population. In this paper we estimate the accuracy of the variance estimates for our current variance estimator, and several design-based modifications and alternatives, using both the jackknife (Wolter 1985) and BRR, through an empirical investigation that more closely reflects the NCS sampling. This is done by first selecting 100 samples of establishments from the our sampling frame, the Longitudinal Database (LDB), using a selection method similar to that used in the NCS. Next, for each variance estimation method under consideration, we calculate 100 variance estimates, one per sample, and then compare the variance estimates with an approximation of the true variance.

Initially, we intended to take 100 independent first-stage samples (with 154 PSUs per sample), and then within each of these samples of PSUs, choose a *single* sample of establishments. Unfortunately, the setup costs of managing a frame and determining sample sizes are high even for a single sample of PSUs, let alone 100 independent samples. Fortunately, for a fixed set of PSUs, the marginal costs of taking multiple samples of establishments are relatively low. Therefore, for this study, we selected 100 independent samples from only one set PSU-sample, namely, the current NCS sample of PSUs.

The true variance is approximated as the sum of a between-PSU component, and a component from the second stage of sampling. For between-PSU variance we use the standard variance formula for a single-stage, one-PSU per stratum design, or a Taylor series approximation in the case of mean wages. The phase-two variance component is computed by taking the mean squared deviation of the 100 sample estimates from their average.

One drawback to the empirical approach considered is that it does not reflect the third stage of sampling, the sampling of occupations. There appears to be no simple way to reflect this stage of sampling in this type of study since the frame contains no occupational information.

Section 2 details the various variance estimators that are studied. The following is the key motivation for these different variance estimators. The sample weights used in this study guarantee that under certain conditions sample estimates $\hat{E}$ of total employment agree with the universe total E for all samples, that is, in keeping with common usage," $\hat{E}$ is calibrated with respect to E," and hence that the variance of $\hat{E}$ is zero.

These conditions are that for each establishment the employment used in calculating the estimates be the same as the frame employment used in selecting the establishment and also the same as the employment used in obtaining a measure of size for each area that is used in the selection of sample areas. (In the NCS, however, these will not be equal, since PSUs were selected using 1995 data, establishments are sampled annually, and once an establishment is sampled, its data is collected for several years.)

All the variance estimators considered involve calculation of a set of replicate estimates based on a set of replicate weights for each replicate. Depending on how these weights are calculated, the replicate estimates of total employment may or may not always be E. If they are always E then the variance estimate of $\hat{E}$ is always zero, in which case the variance estimator of $\hat{E}$ is said to be *zero-calibrated* (Sverchkov, Dorfman, Ernst, and Guciardo 2004). It is desirable for a variance estimator corresponding to $\hat{E}$ to be zero-calibrated if $\hat{E}$ is calibrated with respect to E. We surmised before doing the empirical study that zero-calibrated variance estimators produce more accurate estimates of variance than variance estimators that are not zero-calibrated, with the verification of this hypothesis a key point of this study. Furthermore, among the variance estimators that are not zero-calibrated, we believe that some tend to produce estimates of the variance of $\hat{E}$ that are closer to zero, which is to be preferred if $\hat{E}$ is calibrated with respect to E.

In section 3 we present our empirical results.

2. Jackknife and BRR Variance Estimators Studied

Both the jackknife and BRR are based on the construction of variance strata, with each variance stratum consisting of a set of variance PSUs. These are described in Section 2.1. Next, for both jackknife and BRR, the first step in the calculation of variance estimates is to form a set of replicate estimates in which for each replicate a portion of the sample is either eliminated or down weighted in comparison with the full sample weights used in the basic estimates, and a portion has their sample weights increased. The full sample weights are presented in Section 2.2. Fifteen different approaches to replicate weighting were considered in this study. For jackknife, five basic strategies were used, and for each strategy, we explored two different sets of variance strata: 90 strata and 27 strata. Jackknife methods are presented in Section 2.3. For BRR, an analogous set of five methods were used, yet only for the 90 strata case. BRR methods are presented in Section 2.4. Finally, in Section 2.5, the variance estimation formulas for the jackknife and BRR are presented. From these formulas it is easily seen that that if the replicate estimates corresponding to an estimator must all be the same as the full sample estimate, then the variance estimate must always be zero.

2.1 Variance PSUs and Variance Strata

The variance PSUs are a modification of the actual sample PSUs and each variance stratum is a collection of variance PSUs, with n_g variance PSUs in the *g*-th variance stratum. For the jackknife each actual noncertainty PSU corresponds to a variance PSU and each certainty PSU is split in half to form two variance PSUs as explained in the description of the current variance estimation method in the Introduction. We consider two possibilities for the variance

strata for the jackknife, one uses the same 90 variance strata we use for current BRR estimates in NCS publications, the other uses 27 strata. In the 90 strata case, two variance PSUs, for the most part, are collapsed together to form a variance stratum. In the case of certainty PSUs the corresponding two variance PSUs constitute a variance stratum. For noncertainty PSUs the PSUs collapsed together most be from the same census division, among the nine census divisions, and must be all metropolitan or all nonmetropolitan. For any of these 18 categories for which there are an odd number of variance PSUs, one of the variance strata corresponding to the category consists of three variance PSUs instead of two. In the second set of variance strata for the jackknife there are 27 variance strata corresponding to the 9 census divisions × {certainty metropolitan, noncertainty metropolitan, nonmetropolitan areas}, that is, only one variance stratum per category. We surmised that the second set of variance strata might yield higher variance estimates because generally more variance PSUs that might not as similar are collapsed together to form variance strata with this approach.

For BRR, we considered only one set of variance strata, which is identical to the first set of 90 mentioned for the jackknife and identical to the 90 used in to get NCS published variances. The only difference in the variance PSUs is that for BRR $n_g = 2$ for all strata So for BRR, if a variance stratum has 3 sampled PSUs, then two of these PSUs are combined to form a single variance PSU in order to be able to form variance strata with two variance PSUs in all cases.

2.2 Sample Weighting

Our current weighting system in NCS national wage publications includes factors reflecting the selection probabilities at each of the three stages of selection, and establishment and occupational nonresponse adjustment. It does not presently include any type of calibration or benchmarking, although other survey products in the NCS program do benchmark industry employment estimates to LDB counts, and there are plans to do this for NCS wage estimates.

In our simulation study, the sample weight, w_{igjk}, assigned to establishment k in industry stratum j in variance PSU i within variance stratum g, is the product of the reciprocal of the probability of selecting PSU ig and the reciprocal of the probability of selecting establishment jk conditional on the set of sample PSUs. This is known as the full sample weight to distinguish it from the replicate weights described in Sections 2.3, and 2.4. Weighting factors representing the third stage of sampling and nonresponse adjustment are not used, since there is no third stage of sampling or nonresponse in our study.

We also consider an alternative weighting, where the sample weights just described are adjusted to yield estimates of employment that for each industry stratum j are benchmarked with respect to its frame employment, E_j, since the accuracy of various variance estimation procedures may be different for weights with and without this adjustment (Sverchkov et al. 2004).

2.3 Replicate Weights for the Jackknife

Both the jackknife and BRR variance estimators require that a set of replicate estimates be first calculated, where in the calculation of each replicate estimate a portion of the sample is either eliminated or down weighted, and a portion has their sample weights increased. For both the jackknife and BRR we consider five types of replicate weighting. For both types of variance estimators, the weight w_{igjk} before replicate adjustment is the sample weight without benchmarking to industry employment totals.

In addition, the following modifications to w_{igjk} are used prior to the replicate adjustment. In the case of a certainty establishment in a certainty area, the sample weight is cut in half to compensate for the fact that the establishment is placed in two variance PSUs. In addition, in the BRR case, for those variance strata that consist of three actual PSUs, two of which are collapsed together to form a variance PSU, the sample weights for establishments in the collapsed variance PSU are multiplied by 3/4 and those in the other variance PSU by 3/2 to compensate for the difference in the sum of the weighted employments in these two variance PSUs before this adjustment.

In the case of the jackknife variance estimator considered here, a single variance PSU i' in a single variance stratum g' is deleted in each replicate. (There are other forms of the jackknife, not considered in this paper that delete more than one variance PSU in each replicate.) The five replicate weights, $w^*_{\ell igjk,}\ \ell = 1,...,5$, for establishment $igjk$, for the replicate corresponding to this deletion are as follows:

$$w^*_{1igjk[i'g']} = \begin{cases} 0 & \text{if } i = i', g = g' \\ \dfrac{n_g w_{igjk}}{n_g - 1} & \text{if } i \neq i', g = g', \\ = w_{igjk} & \text{otherwise} \end{cases} \qquad (2.1)$$

where n_g is the number of variance PSUs in variance stratum g.

$w^*_{2igjk[i'g']}$ only differs from $w^*_{1igjki'g'}$ in that

$$w^*_{2ig'jk[i'g']} = \frac{n_{jg'} w_{ig'jk}}{n_{j[i'g']}} \text{ if } i \neq i' \qquad (2.2)$$

where $n_{jg'}$ is the number of sample establishments that are in both industry stratum j and variance stratum g' and $n_{j[i'g']}$ is the number of such establishments that are not in variance PSU $i'g'$.

$w^*_{3igjk[i'g']}$ only differs from $w^*_{1igjki'g'}$ in that

$$w^*_{3igjk[i'g']} = \frac{\hat{E}_{Fjg'}w_{ig'jk}}{\hat{E}_{Fj[i'g']}} \text{ if } i \neq i'$$

where

$$\hat{E}_{Fjg'} = \sum_{ik} w_{ig'jk}E_{Fig'jk} \ , \quad \hat{E}_{Fj[i'g']} = \sum_{i \neq i'}\sum_{k} w_{ig'jk}E_{Fig'jk} \quad \text{and}$$

E_{Figjk} is the frame employment for establishment $igjk$ used in the selection of the establishment.

The rationale for these first three replicate weights is as follows. As noted in the Introduction, $\hat{E}$ is calibrated with respect to E, that is $\sum_{igjk} w_{igjk}E_{igjk} = E$, (where E_{igjk} is the employment of establishment $igjk$ at the time for which estimates are made and E is the universe employment at this time), provided that the employment used to select the sample establishments, the employment used to select the sample areas, and the employment used in estimation are all the same. Consequently, we would want under these conditions for each replicate estimate of total employment to equal the full sample estimate, in which case the estimator of the variance of $\hat{E}$ is zero-calibrated. Now replicate weight 1 does not result in replicate estimates that are equal to the full sample estimate and hence does not produce a zero-calibrated variance estimator. Replicate weight 2 attempts to obtain this zero-calibration by means of adjustment (2.2). It does not completely succeed, however, in part because for each industry stratum, there are different sampling intervals for each of the 54 area sampling clusters, and in part because weighted employment for a certainty establishment differs from that of noncertainty establishments and other certainty establishments in the same sampling cell. Replicate weight 3 does completely succeed in obtaining this zero-calibration. In fact, for replicate weight 3 the variance estimator of total employment is zero even if the establishment sampling is from a different frame than the area sampling, despite the fact that in that case $\hat{E}$ is not calibrated with respect to E and hence the variance estimator with replicate weight 3 tends to underestimate the true variance.

These first three replicate weights are intended for the case when the weights w_{igjk} are not benchmarked with respect to E_j, where E_j is the frame employment of industry j at the time for which estimates are calculated.

The final two replicate weights are for the case when the w_{igjk} are benchmarked with respect to E_j. That is,

$$w^*_{4igjk[i'g']} = \frac{E_j w^*_{1igjk[i'g']}}{\hat{E}_j}$$

$$w^*_{5igjk[i'g']} = \frac{E_j w^*_{1igjk[i'g']}}{\hat{E}_{j[i'g']}}$$

where $\hat{E}_j = \sum_{igk} w_{igjk}E_{igjk}$, $\hat{E}_{j[i'g']} = \sum_{igk} w^*_{1igjk[i'g']}E_{igjk}$. The key difference between these two weights is that for $w^*_{5igjk[i'g']}$, the benchmark factors $E_j / \hat{E}_{j[i'g']}$ are recalculated for the replicate corresponding to each deleted PSU $i'g'$, so that the replicate estimates of employment are calibrated with respect to E_j; while this is not case for $w^*_{4igjk[i'g']}$ since the benchmark factor for each replicate weight is the same as for the corresponding full sample weight.

2.4 Replicate Weights for BRR

For BRR we consider an analogous set of five replicate weighting adjustments to the five considered for the jackknife. While in the jackknife one variance PSU is dropped in the formation of the replicate weights, in BRR one variance PSU is either dropped or down weighted from each variance stratum, and the sample establishments in the other variance PSU in each variance stratum has its weights increased.

For BRR, we let S_r denote, the set of variance PSUs that have their weights increased in the r-th replicate and S'_r denote the complimentary set. The S_r are determined by specifying a Hadamard matrix as described in Wolter (1985).

We let f denote a factor applied to variance PSUs in S_r, with $f' = 2 - f$ the corresponding factor in S'_r. For standard BRR, $f = 2$ and hence the variance PSUs in S'_r are dropped. Fay's method corresponds to $1 < f < 2$ with $f = 1.5$ the most commonly used value for Fay's method,. For Fay's method, all units receive positive weights.

The five replicate weights, $w^*_{\ell igjkr,}$ $\ell = 1,...,5$, for BRR, analogous to the replicate weights for the jackknife, with the subscript r denoting the replicate r, are

$$w^*_{1igjkr} = \begin{cases} fw_{igjk} \text{ if } ig \in S_r \\ f'w_{igjk} \text{ if } ig \in S'_r \end{cases} \quad (2.3)$$

$$w^*_{2igjkr} = \frac{n_j w^*_{1igjkr}}{fn_{jr} + f'n'_{jr}}$$

where n_{jr}, n'_{jr} and n_j are the total number of sample establishments in S_r, S'_r, and in all variance PSUs, respectively, in industry j.

$$w^*_{3igjk} = \frac{\hat{E}_{Fj}w^*_{1igjkr}}{\hat{E}_{Fjr}}$$

where $\hat{E}_{Fj} = \sum_{igk} w_{igjk} E_{Figjk}$, $\hat{E}_{Fjr} = \sum_{igk} w^*_{1igjkr} E_{Figjk}$.

Note that, unlike in the case of the jackknife, replicate weight 3 does not completely restore the calibration with respect to E. This is solely due to variance strata with three actual PSUs, and the weighting factors of 3/4 and 3/2 that are used on the corresponding variance PSUs, as explained in Section 2.3, and we anticipated that the impact of this problem would be small.

The fourth and fifth replicate weights are completely analogous to that for the jackknife. That is,

$$w^*_{4igjkr} = \frac{E_j w^*_{1igjkr}}{\hat{E}_j}$$

$$w^*_{5igjkr} = \frac{E_j w^*_{1igjkr}}{\hat{E}_{jr}}$$

where $\hat{E}_{jr} = \sum_{igk} w^*_{1igjkr} E_{igjk}$.

2.5 Variance Estimation Formula for BRR and the Jackknife

For the jackknife and BRR, the replicate estimates are obtained using the same form of estimator as the full sample estimator, with full sample weights replaced by the replicate weights described in the previous two subsections.

For the jackknife, with $\hat{\Theta}$ the full sample estimator and $\hat{\Theta}_{ig}$ the corresponding replicate estimator with PSU ig deleted, the jackknife variance estimator v_J is

$$v_J = \sum_g \left(\frac{n_g - 1}{n_g} \right) \sum_i (\hat{\Theta}_{[ig]} - \hat{\Theta})^2$$

For the BRR with factor f and with $\hat{\Theta}_{rf}$ the replicate estimate corresponding to replicate r and factor f, the BRR variance estimator with factor f, denoted v_{BRRf}, is

$$v_{BRRf} = \frac{1}{R(1-f)^2} \sum_r (\hat{\Theta}_{rf} - \hat{\Theta})^2$$

where R *is* the number of replicates.

3. Empirical Results

The empirical study began with selection of 100 samples of establishments from our LDB sampling frame. We generally followed the current method of selection for the NCS wage sample as described in Ernst, Guciardo, Ponikowski, and Tehonica (2002). However, while under the current method most of the sample is obtained from a rotating panel design, with each annual sample panel consisting of one fifth of the entire rotating sample selected from an updated frame, we selected all five sample panels from a single frame to simplify our work. We excluded from our sampling frame and our estimates all government and aerospace manufacturing establishments. This is because, although such establishments do contribute to our current production estimates, they were selected from an

older and somewhat different sample design. Also, in addition to calculating variance estimates for the entire nation (All U.S.), we calculated variance estimates for the domain "remaining 99 primary sampling strata" (Rem 99) corresponding to the cluster of the remaining 99 PSUs. This cluster is of greatest concern, since the sampling is not independent in each variance PSU corresponding to this cluster. (We did not study the three-PSU Alaska-Hawaii cluster, since this cluster is very small.). In our weight adjustments, two definitions for industry j were used. For All U.S. and its subdomains, the j were the 23 industry strata within All U.S. For Rem 99 and its subdomains, the j were the same 23 industry strata, yet restricted to Rem 99. This ensures that Rem 99 total employment estimates are calibrated, as well as the All U.S. estimates.

The results are presented in 10 different tables, covering all combinations of domains (All U.S., Rem 99, and their 23 industry strata), estimates (total employment and mean quarterly wages), and variance estimation method (jackknife with 90 variance strata, jackknife with 27 variance strata, and Fay's BRR with $f = 1.5$). Let the two jackknife methods be known as jackknife-90 and jackknife-27. Tables 1 and 2 display detailed results for the two domains All U.S. and Rem 99. Tables 4-10 have summary results for the 23 industries in both All U.S. and Rem 99. Tables 4-7 compare standard errors across the various methods. Tables 7-10 provide summary information on 95% confidence interval coverage rates.

In Tables 1 and 2, the rows correspond to variance estimators, for All U.S. or Rem. 99. The first column is an estimate $\hat{\sigma}$ of the true standard error for the non-benchmarked case. The next 9 columns can be divided into three groups of three with each group corresponding to one of the sets of replicate weights $w^*_{\ell igjk[i'j']}$, $\ell = 1,2,3$, in the case of the jackknife, and $w^*_{\ell igjkr}$, $\ell = 1,2,3$, in the case of BRR. The first column in each group is a ratio equal to the "root average variance estimate" (RAVE), divided by the estimate $\hat{\sigma}$ of the true standard error in column 1. The RAVE is the square root of the arithmetic mean over the 100 samples of the variance estimate $\hat{V}(i)$, $i = 1,...,100$ (calculated using the variance estimation formula corresponding to the group). Values near 1 are to be preferred in this column. The second column in a group is the percentage of the 100 samples for which the estimated 95% confidence interval covers the true value. The final column in each group is an estimate of the "percent relative root mean squared error":

$$\% \, Rel. \, RMSE = 100 \sqrt{ \tfrac{1}{100} \sum_{i=1}^{100} \left[\sqrt{\hat{V}(i)} - \hat{\sigma} \right]^2 } \Big/ \hat{\sigma}$$

which is the square root of the mean squared error of the 100 estimates of standard error, expressed as a percent of the estimated true standard error.

The remaining columns will display output for the benchmarked cases ($\ell = 4,5$). For mean wages, the format is identical. For total employment, however, we will only show the square root of the average variance estimate for $\ell = 4$. All other results are trivial, because for both methods, $\hat{E}$ is calibrated with respect to E, and for $\ell = 5$, all variance estimators are zero-calibrated. The former property ensures that the true variance is zero, and that each confidence interval contains E (we are using closed intervals, which contain $\hat{E} = E$, even when the variance estimate is zero). The latter property ensures that, for $\ell = 5$, all standard error estimates equal zero, in which case the standard error of the standard error estimator is zero.

Table 1 shows output for total employment, for All U.S and Rem 99. The square roots of the average variance estimates (or RAVE values) decrease from method 1 to 2 to 3. Values for method 1 are above the estimated true standard error, method 2 is above or very close, and method 3 is substantially smaller, which all was expected (see Section 2.3). Methods 1 and 2 have all of their confidence intervals containing frame employment, yet method 3 has less coverage. The estimated MSEs of the standard error estimators are relatively high for most methods, yet this is more because of the contribution to the MSE of the bias squared term then than the variance term. The bias squared

term of this MSE was estimated by $(\hat{\alpha} - \hat{\sigma})^2$ and the variance term by $\frac{1}{100} \sum_{i=1}^{100} \left[\sqrt{\hat{V}(i)} - \hat{\alpha} \right]^2$, where $\hat{\alpha} = \frac{1}{100} \sum_{i=1}^{100} \sqrt{\hat{V}(i)}$. These terms are omitted from the tables. Fay's BRR for method 2 has the least bias in the standard error estimator. For methods 1 and 2, Jackknife-90 have RAVE values similar to Fay's BRR, which was expected, because they use the same 90 variance strata. Jackknife-27 has substantially higher RAVE values for methods 1 and 2, which was expected, because it has fewer variance strata and hence a larger collapsed stratum effect.

In Table 2, we have output for mean wages for All U.S. and Rem 99. The RAVE values are above the estimated true standard error. For All U.S., as expected, the RAVE values drop from methods 1 to 2 to 3, and method 5 (where we rebenchmark for each replicate) results in lower RAVE values than method 4. Some results for Rem 99 were unexpected, since RAVE values for both jackknife methods increase from method 1 to 2 to 3, and from method 4 to 5. Overall, Fay's BRR has RAVE values closest to the estimated true value, particularly for methods 2 and 5, and confidence interval coverage is closer to 95%, in general.

Table 1. Output for Total Employment											
	Non-Benchmarked Cases										**Bench**
Domain, Variance Estimator	**True**	**Method L = 1**			**Method L = 2**			**Method L = 3**			**L = 4**
	$\hat{\sigma}$ / 1000	$\sqrt{Avg(\hat{V})}$ / $\hat{\sigma}$	Conf Inter Cov %	% Rel Rmse of $\sqrt{\hat{V}}$	$\sqrt{Avg(\hat{V})}$ / $\hat{\sigma}$	Conf Inter Cov %	% Rel Rmse of $\sqrt{\hat{V}}$	$\sqrt{Avg(\hat{V})}$ / $\hat{\sigma}$	Conf Inter Cov %	% Rel Rmse of $\sqrt{\hat{V}}$	$\sqrt{Avg(\hat{V})}$ / 1000
All US, Jack-90	688	2.08	100	108	1.41	100	41	0.33	100	67	1383
All US, Jack-27	688	9.09	100	809	6.73	100	573	0.17	83	83	6241
All US, Fay's	688	1.90	100	90	1.00	100	3	0.13	62	87	1278
Rem 99, Jack-90	659	1.85	100	86	1.38	100	39	0.33	100	67	1095
Rem 99, Jack-27	659	3.57	100	257	1.44	100	44	0.18	99	82	2292
Rem 99, Fay's	659	1.77	100	77	1.00	100	3	0.10	68	91	1095

Table 2. Output for Mean Quarterly Earnings																
	Non-Benchmarked Cases										**Benchmarked Cases**					
Domain, Variance Estimator	**True**	**Method L = 1**			**Method L = 2**			**Method L = 3**			**True**	**Method L = 4**			**Method L = 5**	
	$\hat{\sigma}$	$\sqrt{Avg(\hat{V})}$ / $\hat{\sigma}$	Conf Inter Cov %	% Rel Rmse of $\sqrt{\hat{V}}$	$\sqrt{Avg(\hat{V})}$ / $\hat{\sigma}$	Conf Inter Cov %	% Rel Rmse of $\sqrt{\hat{V}}$	$\sqrt{Avg(\hat{V})}$ / $\hat{\sigma}$	Conf Inter Cov %	% Rel Rmse of $\sqrt{\hat{V}}$	$\hat{\sigma}$	$\sqrt{Avg(\hat{V})}$ / $\hat{\sigma}$	Conf Inter Cov %	% Rel Rmse of $\sqrt{\hat{V}}$	$\sqrt{Avg(\hat{V})}$ / $\hat{\sigma}$	Conf Inter Cov %	% Rel Rmse of $\sqrt{\hat{V}}$
All US, Jack-90	47	1.33	100	35	1.25	98	29	1.31	100	34	47	1.31	100	34	1.25	100	29
All US, Jack-27	47	3.82	100	282	2.79	100	179	2.38	100	138	47	3.80	100	280	3.27	100	227
All US, Fay's	47	1.28	100	32	1.17	96	23	1.13	96	22	47	1.27	99	31	1.12	99	21
Rem 99, Jack-90	75	1.52	100	53	1.67	100	69	1.87	100	88	78	1.49	100	50	1.68	100	69
Rem 99, Jack-27	75	1.66	100	67	1.79	100	80	2.08	100	108	78	1.63	100	64	1.80	100	81
Rem 99, Fay's	75	1.50	100	53	1.39	100	41	1.45	100	47	78	1.47	100	50	1.37	100	40

Note: In the table headers above, the columns repeat for each method — $\sqrt{Avg(\hat{V})}/\hat{\sigma}$ (with $\sqrt{Avg(\hat{V})}$ over $\hat{\sigma}$), Conf Inter Cov %, and % Rel Rmse of $\sqrt{\hat{V}}$.

Tables 3-8 record, for each cell, the percentage of the 23 industry domains where the RAVE value of the method indicated by the column exceeds the RAVE value of the method indicated by the row. The first row and column refers to the estimated true standard error for the non-benchmarked cases.

In Table 3 (for total employment, All US industries), Jackknife-27 for $\ell=1,2$ dominates (that is, the percentages in the corresponding columns are high) and method 3 in most cases is dominated by methods 1 and 2. Fay's BRR with $\ell=2$ has a RAVE value greater than the true standard error estimate for 43% of domains, which is closest to 50% of all methods. (Note that a dominance of about 50% over the estimated true standard error seems desirable as supporting unbiasedness of the variance estimator.) In Table 4, (for total employment, Rem 99 industries), jackknife 27 is less dominant. Both jackknife-27 and jackknife-90 with $\ell=2$ have 52% dominance over the estimated true variance, the closest to 50%.

Table 3. Comparing Average Standard Errors
Total Employment, 23 Industries in All US

Percent of 23 Domains where Col > Row	True	Jack-90 (L=1)	Jack-90 (L=2)	Jack-90 (L=3)	Jack-27 (L=1)	Jack-27 (L=2)	Jack-27 (L=3)	Fay's (L=1)	Fay's (L=2)	Fay's (L=3)
True	–	96	74	13	100	100	9	96	43	0
Jack-90 (L=1)	4	–	30	0	96	91	0	57	9	0
Jack-90 (L=2)	26	70	–	0	91	96	0	74	9	0
Jack-90 (L=3)	87	100	100	–	100	100	4	100	100	9
Jack-27 (L=1)	0	4	9	0	–	22	0	4	0	0
Jack-27 (L=2)	0	9	4	0	78	–	0	4	0	0
Jack-27 (L=3)	91	100	100	96	100	100	–	100	100	4
Fay's (L=1)	4	43	26	0	96	96	0	–	4	0
Fay's (L=2)	57	91	91	0	100	100	0	96	–	0
Fay's (L=3)	100	100	100	91	100	100	96	100	100	–

Table 4. Comparing Average Standard Errors
Total Employment, 23 Industries in Rem 99

Percent of 23 Domains where Col > Row	True	Jack-90 (L=1)	Jack-90 (L=2)	Jack-90 (L=3)	Jack-27 (L=1)	Jack-27 (L=2)	Jack-27 (L=3)	Fay's (L=1)	Fay's (L=2)	Fay's (L=3)
True	–	78	52	13	96	52	13	87	39	0
Jack-90 (L=1)	22	–	26	0	96	30	0	78	17	0
Jack-90 (L=2)	48	74	–	0	78	43	9	70	35	0
Jack-90 (L=3)	87	100	91	–	100	96	17	100	91	9
Jack-27 (L=1)	4	4	22	0	–	26	0	22	13	0
Jack-27 (L=2)	48	70	57	4	74	–	0	61	30	0
Jack-27 (L=3)	87	100	91	83	100	91	–	100	91	9
Fay's (L=1)	13	22	30	0	78	39	0	–	17	0
Fay's (L=2)	61	83	65	9	87	70	9	83	–	0
Fay's (L=3)	100	100	100	91	100	100	91	100	91	–

For total employment in each industry j for benchmarked cases $\ell=4,5$, the situation is the same as it is for total employment in all industries combined; that is, for both methods $\hat{E}$ is calibrated with respect to E, and for $\ell=5$, all variance estimators are zero-calibrated.

In Tables 5 and 6, for mean wages, jackknife 27 with $\ell=1,2$ dominates, although less so for industries in Rem 99. All the values in row 1 of each table are greater than 50 %, with Jackknife-90 with $\ell=3$ having the values closest to 50%. For mean wages, for an industry j, the Fay's BRR variance estimates for $\ell=2,3$ are equivalent to those for $\ell=1$ because the replicate adjustment factors (multiplied by w_{igjk}) are the same for all units in j, in both the numerator and denominator of the mean wage estimator, and hence cancel out leaving only the formula for $\ell=1$.

For mean wages in each industry j for benchmarked cases $\ell=4,5$, variance estimates are the same as those for the case $\ell=1$, because each industry j is a benchmark cell. Therefore, all units have the same benchmark adjustment factor in the replicate estimates, and the factor cancels out.

Table 5. Comparing Average Standard Errors
Mean Wage, 23 Industries in All U.S.

Percent of 23 Domains where Col > Row	True	Jack-90 (L=1)	Jack-90 (L=2)	Jack-90 (L=3)	Jack-27 (L=1)	Jack-27 (L=2)	Jack-27 (L=3)	Fay's (L=1,2,3)
True	–	78	87	65	96	96	96	78
Jack-90 (L=1)	22	–	74	35	100	100	100	35
Jack-90 (L=2)	13	26	–	13	87	96	78	26
Jack-90 (L=3)	35	65	87	–	96	96	100	61
Jack-27 (L=1)	4	0	13	4	–	48	22	4
Jack-27 (L=2)	4	0	4	4	52	–	26	4
Jack-27 (L=3)	4	0	22	0	78	74	–	9
Fay's (L=1,2,3)	22	65	74	39	96	96	91	–

Table 6. Comparing Average Standard Errors
Mean Wage, 23 Industries in Rem 99

Percent of 23 Domains where Col > Row	True	Jack-90 (L=1)	Jack-90 (L=2)	Jack-90 (L=3)	Jack-27 (L=1)	Jack-27 (L=2)	Jack-27 (L=3)	Fay's (L=1,2,3)
True	–	83	83	61	91	100	91	70
Jack-90 (L=1)	17	–	65	39	83	91	87	35
Jack-90 (L=2)	17	35	–	17	39	65	52	30
Jack-90 (L=3)	39	61	74	–	65	91	96	61
Jack-27 (L=1)	9	17	61	35	–	78	70	30
Jack-27 (L=2)	0	9	35	9	22	–	13	17
Jack-27 (L=3)	9	13	48	4	30	78	–	22
Fay's (L=1,2,3)	30	65	70	39	70	83	78	–

Tables 7-10 summarize confidence interval coverage rates; that is, the percent of the samples where the estimated 95% confidence interval contains the true employment or wage. Each table-cell displays the percentage of the 23 industries that have coverage rates X within the given range

Tables 7 and 8 clearly show how extreme the confidence interval coverage can be for employment. Most rates are high, but rates are low for $\ell = 3$ since variance estimates are too close to zero.

Table 7. Confidence Interval Coverage Rates
Total Employment, 23 Industries in All US

Percent of 23 Domains where 95% Confid Interv Coverage X is:	Jack-90 (L=1)	Jack-90 (L=2)	Jack-90 (L=3)	Jack-27 (L=1)	Jack-27 (L=2)	Jack-27 (L=3)	Fay's (L=1)	Fay's (L=2)	Fay's (L=3)
X ≥ 98%	96	96	39	100	100	13	96	87	4
90% ≤ X < 98%	4	4	13	0	0	9	4	4	9
80% ≤ X < 90%	0	0	0	0	0	9	0	0	9
X < 80%	0	0	48	0	0	70	0	9	78

Table 8. Confidence Interval Coverage Rates
Total Employment, 23 Industries in Rem 99

Percent of 23 Domains where 95% Confid Interv Coverage X is:	Jack-90 (L=1)	Jack-90 (L=2)	Jack-90 (L=3)	Jack-27 (L=1)	Jack-27 (L=2)	Jack-27 (L=3)	Fay's (L=1)	Fay's (L=2)	Fay's (L=3)
X ≥ 98%	96	87	39	96	87	17	96	83	4
90% ≤ X < 98%	0	4	17	0	0	4	0	0	9
80% ≤ X < 90%	0	0	0	0	4	4	0	0	0
X < 80%	4	9	43	4	9	74	4	17	87

Tables 9 and 10 for mean wages show less extreme coverage rates, yet coverage is usually too high. Jackknife-90 for $\ell = 3$ and Fay's BRR have the least amount of overcoverage.. Also in Table 10, there is some evidence of undercoverage.

Table 9. Confidence Interval Coverage Rates
Mean Wage, 23 Industries in All US

Percent of 23 Domains where 95% Confid Interv Coverage X is:	Jack-90 (L=1)	Jack-90 (L=2)	Jack-90 (L=3)	Jack-27 (L=1)	Jack-27 (L=2)	Jack-27 (L=3)	Fay's (L=1,2,3)
X ≥ 98%	48	52	39	78	78	70	43
90% ≤ X < 98%	43	39	52	13	13	22	48
80% ≤ X < 90%	4	4	4	4	4	4	4
X < 80%	4	4	4	4	4	4	4

Table 10. Confidence Interval Coverage Rates
Mean Wage, 23 Industries in Rem 99

Percent of 23 Domains where 95% Confid Interv Coverage X is:	Jack-90 (L=1)	Jack-90 (L=2)	Jack-90 (L=3)	Jack-27 (L=1)	Jack-27 (L=2)	Jack-27 (L=3)	Fay's (L=1,2,3)
X ≥ 98%	52	57	48	57	65	52	57
90% ≤ X < 98%	26	26	30	22	22	35	22
80% ≤ X < 90%	9	4	9	9	4	4	9
X < 80%	13	13	13	13	9	9	13

4. Conclusions

For total employment, we found method 2, particularly for Fay's BRR, to be the class of standard error estimators with the least bias for most key domains when the estimates are not benchmarked. Method 1 is generally too high, method 2 is better, method 3 is too low. This is consistent with what we surmised when we introduced these three methods in Section 2. Method 5 tends to do better than method 4, since it re-benchmarks for each replicate.

For mean wages, these trends are less prominent, and in some situations reversed, particularly for jackknife-90 and the Rem 99 domains. Yet Fay's BRR does well in many cases, especially for method 2.

5. References

Ernst, L. R., Guciardo, C. J., Ponikowski, C. H., and Tehonica, J. (2002). Sample Allocation and Selection for the National Compensation Survey. 2002 Proceedings of the American Statistical Association, Section on Survey Research Methods, [CD-ROM], Alexandria, VA: American Statistical Association.

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

Sverchkov, M., Dorfman, A. H., Ernst, L. R., Guciardo, C. J., and (2004). Zero-Calibrated Variance Estimators. Proceedings of the American Statistical Association, Section on Survey Research Methods, [CD-ROM].

Wang, S., Dorfman, A. H., and Ernst, L. R. (2004). A New Variance Estimator for a Two-phase Restratified Sample with PPS at Both Phases. Journal of Statistical Planning and Inference, (to appear).

Wolter, K. M (1985). Introduction to Variance Estimation. New York: Springer-Verlag.