Implementation of Controlled Selection in the National Compensation Survey Redesign October 2008

Lawrence R. Ernst [1], Christopher J. Guciardo [2], Yoel Izsak [3], Jonathan J. Lisic [2], Chester H. Ponikowski[2]

[1] Bureau of Labor Statistics, 2 Massachusetts Ave., N.E., Room 1950, Washington, DC 20212, Ernst.Lawrence@bls.gov

[2] Bureau of Labor Statistics, 2 Massachusetts Ave., N.E., Room 3160, Washington, DC 20212

[3] National Agricultural Statistics Service, 1400 Independence Ave., S.W., Room 6337, Washington, DC 20250

Abstract

Izsak et al. (2005) included a proposal to allocate the number of sample establishments among the sampling cells using a controlled selection procedure, where cells are area PSUs × industry sampling strata × sampling panels. Since then the procedure has been implemented but with a number of modifications not discussed in Izsak et al. (2005). These modifications and possible future changes are discussed in this new paper. They include: weighting changes necessitated by the use of controlled selection, complications caused by rounding issues and how they were overcome, complexities caused by the need to allocate over five sampling panels, and use of a real-valued minimum allocation for each sampling cell in the controlled selection process in order to avoid very large sample weights and accompanying increases in variances.

Key Words: Allocation, sampling panels, rounding, sampling cells, area PSUs, industry sampling strata

1. Introduction

This paper covers some issues regarding the most recent redesign of the National Compensation Survey (NCS), a compensation program conducted by the Bureau of Labor Statistics, that were not completely covered in previous papers, such as Izsak et al. (2005). In particular, this paper covers some of the aspects of the use of controlled selection in the sample allocation process. This paper is best read in conjunction with Izsak et al. (2005), and we will not repeat material that overlaps with that paper, to the extent that seems reasonable. Other papers that discuss some aspects of the NCS sample redesign include Ernst et al. (2002), Ernst, Guciardo, and Izsak (2004), Ernst, Izsak, and Paben (2004), and Izsak et al. (2003).

A key step in the selection of sample establishments for the NCS is the allocation of the NCS Wage Only sample and the Employment Cost Index (ECI) among 152 area strata × 20 industry strata for the government sector and 152 area strata × 23 industry strata × five sample panels for the private sector. The allocations are determined by solving several controlled selection problems using the methodology described in Causey, Cox, and Ernst (1985). The use of this methodology for the NCS application was first proposed in Izsak et al. (2005). However, the controlled selection procedure actually implemented required a number of modifications that were not discussed in that paper. They include: weighting changes necessitated by the use of controlled selection, complications caused by rounding issues and how they were overcome, complexities caused by the need to allocate over five sampling panels, and use of a real-valued minimum allocation for each sampling cell in the controlled selection process in order to avoid very large sample weights and accompanying increases in variances.

A short description of the two-dimensional controlled selection problem in general is presented in Section 2. The specific formulations of the controlled selection problems used in the NCS application are discussed in Section 3. Some necessary modifications to insure that cell values for the internal cells of the controlled selection problem sum to the necessary marginals and that this additivity is not destroyed by either rounding or the fact that the original controlled selection arrays are generally real-valued, not integer-valued, are presented in Section 4. Weighting changes necessitated by the use of controlled selection are presented in Section 5. Finally, in Section 6, the use of real-valued minimum allocations for the sampling cells in the controlled selection process in order to avoid very large sample weights and accompanying increases in variances is discussed along with other options for modifying the allocation process.

2. Controlled Selection Problems

Much of this paper is concerned with obtaining allocations of portions of the NCS sample among sample area × industry stratum cells by construction of two-dimensional tabular, that is additive, arrays, each of which constitutes a controlled selection problem, and then solving each controlled selection problem using a modification of the method of Causey, Cox, and Ernst (1985).

Note that a two-dimensional controlled selection problem in the context of this paper is a two-dimensional additive array

$\mathbf{S} = (s_{ij})$ of dimensions $(M+1)\times(N+1)$, where M is the number of sample areas and N is the number of industry strata. $\mathbf{S}$ satisfies the following conditions: Each internal cell value, that is, a cell value for a cell that is neither in the last row or last column, is the expected number of sample units to be selected in the corresponding sample area × industry stratum cell. A row marginal is the value for a cell in the last column of a row and a column marginal is the value for a cell in the last row of a column. These two marginals are, respectively, the expected number of sample units in an area and an industry. The cell value in the final row and column is the total sample size. The cell values, except in some cases for the grand total, are generally real-valued, not integer-valued. A solution to the controlled selection problem $\mathbf{S}$ is a set of ℓ integer-valued two-dimensional additive arrays, $\mathbf{N}_1 = (n_{ij1}),...,\mathbf{N}_\ell = (n_{ij\ell})$, of the same dimensions as $\mathbf{S}$, and associated probabilities, $p_1,..., p_\ell$, such that for each cell *ij*, including marginals, in each array $\mathbf{N}_k$

$$\left| n_{ijk} - s_{ij} \right| < 1 \tag{2.1}$$

and for each *ij*

$$\sum_{k=1}^{\ell} p_k n_{ijk} = s_{ij} \tag{2.2}$$

A solution to a controlled selection problem can be obtained by solving a sequence of transportation problems through a recursive procedure described in Causey, Cox, and Ernst (1985 Sec. 4.1). Once a solution is obtained to a controlled selection problem, one of the arrays $\mathbf{N}_k$ is chosen from among the arrays $\mathbf{N}_1,...,\mathbf{N}_\ell$ using the associated probabilities $p_1,..., p_\ell$. The chosen $\mathbf{N}_k$ determines the allocation to each sampling cell. Note that by (2.1) and (2.2) it follows that for the solution to our controlled selection problems:

The number of sample units in each sample area, in each industry stratum, and in each sample area × industry stratum cell is within one of the desired number for every possible sample. (2.3)

The expected number of sample units in each of the domains listed in (2.3) over all possible samples is the desired number. (2.4)

However, the methodology given in Causey, Cox, and Ernst (1985) for solving a controlled selection may fail to yield a solution to a controlled selection problem due to rounding error in the computation of the cell values, since the rounding error can destroy the additivity of the tabular array. In particular, this can occur for controlled selection problems associated with the NCS allocations to be described, since the cell values for each of the initial controlled selection arrays are real numbers as opposed to integers.

We will explain in Section 4 how the problems arising from rounding error can be avoided for these controlled selection problems by converting the original controlled selection problem into a sequence of rounding problems involving integer arithmetic. Before doing so, however, we proceed to describe in Section 3 how the original real-valued tabular arrays were constructed for our NCS allocation problems.

3. Formulation of the Tabular Arrays for NCS Allocations

There are five types of tabular arrays that require the solution of controlled selection problems in conjunction with the NCS allocations. We first consider the arrays for the government sector. For this sector there are two controlled selection problems, one for the ECI allocation and one for the NCS Wage Only allocation, with the latter problem created by subtracting the ECI allocation array from the total NCS Wage allocation array. Note, in particular, that although a controlled selection problem for total government NCS Wage is constructed, it is not solved directly. For reasons explained in Izsak et al. (2005), the Wage Only array is solved instead and the sum of the solution to the Wage Only array and the ECI allocation array is used as a solution to the total NCS Wage allocation array.

The preliminary government ECI allocation array is obtained by taking the entire ECI government sample count, which is an integer, and allocating among the 20 government sampling industries proportional to PSU weighted employment, with the industry allocations being real-valued with no rounding. The allocation for each of the 20 ECI government industry totals are then allocated among 152 sample areas, again proportional to PSU weighted employment. This creates a preliminary two-dimensional, real-valued controlled selection problem with dimensions of the internal cells being 152 × 20.

The preliminary total NCS Wage government allocation is obtained similarly to the ECI array, but in the opposite order. The total government NCS Wage sample is first allocated among the 152 areas and then within each area to the 20 industries. The allocation arrays for government ECI and total government NCS wage are then modified in two ways. First, for any cell in which the total NCS Wage allocation is less than the ECI allocation, the NCS Wage allocation is raised to the ECI allocation. In addition, any cell for which either the ECI or the total NCS Wage allocation exceeds the number of frame units has its cell allocation

lowered to the number of frame units for that survey or surveys. Then, for each of these surveys, the remaining sample in each area is reallocated among the remaining cells in the area proportional to frame employment and the allocation adjustment process iterated until no more adjustments are necessary. The modified marginals are then obtained by summing. Finally, as stated earlier in this section, the controlled selection array for ECI government is subtracted from the controlled selection array for total NCS Wage government to obtain a controlled selection array for NCS Wage Only government and the controlled selection procedure is used to obtain integer allocations for ECI and NCS Wage Only government.

For the private sector there are three types of controlled selection problems used in the allocation process. First, the total NCS Wage private sector sample and ECI sample over five panels (Izsak et al. 2005) is allocated among the sampling cells. This is analogous to the government sector except there are 23 industries in the private sector. Also, when allocating the private sector ECI sample among the industries, the sample allocation to each industry is based on historical allocations that in turn are based on frame employments, variances, and response rates for the industries, instead of allocating proportional to PSU weighted employment. A topic for further research is possible adjustments of the proportions of the ECI allocated to each industry to take into account changes in these quantities by industry over time.

An additional difference for the private sector is that for the NCS Wage sample there is a minimum allocation for the Pay Agent areas (Izsak et al. 2005)) and a maximum allocation for the three largest areas. When allocating this sample among the areas, any area with an allocation below the minimum or above the maximum allocation has its allocation adjusted to the minimum or maximum, respectively, with the remaining sample allocated to the remaining areas proportional to PSU weighted employment and the allocation adjustment process iterated if necessary. The total private sector sample for ECI and NCS Wage are integer-valued, but the cell allocations are real-valued for all other cells. Controlled selection is not used to directly allocate either the total private NCS Wage Only sample or the ECI sample over five panels. The first private sector controlled selection array is for the ECI five panel noncertainty units as explained in Izsak et al. (2005) and summarized here. We first obtain the allocation to the ECI noncertainty five panel sample in each cell by subtracting the number of five panel ECI certainties from the total ECI five panel sample size in the cell, and allocating the remaining ECI sample between five panel noncertainties and single panel units, proportional to PSU weighted employment. The marginals for five panel ECI noncertainties are obtained by summing the internal cells. Controlled selection is then performed on the tabular array for ECI five panel noncertainties. The grand total for the ECI five panel noncertainties and the single panel units are integer-valued, but the cell allocations are real-valued for all other cells.

The final two controlled selection problems are for single panel private sector noncertainty units for ECI and NCS Wage Only. These are done similarly to the two controlled selection problems for the government sector, with the following exceptions. The ECI single panel cell allocation for the first of the five panels is obtained by taking the total ECI allocation for the cell, subtracting the number of ECI five panel sample units, and dividing by 5. The NCS Wage single panel cell allocation for the first of the five panels is handled similarly. The NCS Wage Only allocation for this panel is obtained by subtraction as was done for the government sector.

Originally it was intended that the resulting controlled selection arrays for the ECI and NCS Wage Only single panels would each be used to independently obtain five controlled roundings corresponding to the five single panels for ECI and NCS Wage Only. However, for the first single panel sample, a sample cut took place before the sample selection. The result of the sample cut was a reduction of the allocation in each ECI sampling cell in the ECI controlled selection array by a fixed percentage, and a reduction in each NCS Wage sampling cell in the NCS Wage controlled selection array by a different fixed percentage. These sample reductions were performed with the constraint that the NCS Wage sample for each Pay Agent area not be reduced below the Pay Agent area minimum. For the second single panel sample, an additional sample cut took place and the cell allocations were also modified by the use of updated frame counts for the sampling cells, with these two changes resulting in a different controlled selection array. where this cut was based on the reduced expected allocation in each cell after the cut for the first single panel sample. Thus the controlled selection problems differed between the first and second single panels and will continue to differ if there are further sample adjustments corresponding to the other single panel samples or if updated frames are used in forming the controlled selection arrays.

Note that the altered cell allocations used in the controlled problems for the single panel samples are generally real-valued, not integer-valued, even for the grand total because of the division by 5.

4. Modifying Controlled Selection Problems to Avoid Rounding Errors

In this section we explain how modifications are made to avoid rounding errors that would otherwise destroy the additivity of the controlled selection arrays. These modifications involve conversion of the initial controlled selection array. Note in general for a two-dimensional tabular array $\mathbf{S} = (s_{ij})$, a controlled rounding of $\mathbf{S}$ to a positive integer base b is a tabular array $\mathbf{N} = (n_{ij})$, where for each ij, n_{ij} is a positive integer multiple of b for which $|n_{ij} - s_{ij}| < b$. If no base is specified, then base 1 is understood, that is $\mathbf{N} = (n_{ij})$ is an integer-valued tabular array satisfying (2.1).

For each of the controlled selection problems $\mathbf{S} = (s_{ij})$ described in Section 3, the array **S** is generally not integer-valued, which can lead to lack of additivity and inability to solve the necessary transportation problems to obtain controlled roundings of **S**. To convert the array to a controlled selection array which overcomes these difficulties, we convert the array **S** with dimensions $(M+1)\times(N+1)$ to an integer-valued additive array $\mathbf{S}' = (s'_{ij})$ with dimensions $(M+2)\times(N+2)$, incorporating an approach in Cox and Ernst (1982). The marginals of $\mathbf{S}'$ are positive integer multiples of a base 10^{γ}, where γ is a positive integer that depends on the number of places of accuracy desired. $\gamma = 4$ was used for the controlled selection problems considered for NCS. To obtain $\mathbf{S}'$ first let

$$s'_{ij} = \text{floor}(10^{\gamma} s_{ij}, 1), i = 1,..., M,\ j = 1,..., N, \tag{4.1}$$

$$s'_{i(N+2)} = \text{ceiling}(\textstyle\sum_{j=1}^{N} s'_{ij}, 10^{\gamma}), i = 1,..., M, \tag{4.2}$$

$$s'_{i(N+1)} = s'_{i(N+2)} - \textstyle\sum_{j=1}^{N} s'_{ij}, i = 1,..., M, \tag{4.3}$$

$$s'_{(M+2)j} = \text{ceiling}(\textstyle\sum_{i=1}^{M} s'_{ij}, 10^{\gamma}),\ j = 1,..., N+1, \tag{4.4}$$

$$s'_{(M+1)j} = s'_{(M+2)j} - \textstyle\sum_{i=1}^{M} s'_{ij},\ j = 1,..., N+1, \tag{4.5}$$

$$s'_{i(N+2)} = \textstyle\sum_{j=1}^{N+1} s'_{ij}, i = M+1, M+2, \tag{4.6}$$

where floor (x, y) is the largest integer multiple of y not exceeding x and ceiling (x, y) is the smallest integer multiple of y that is not less than x.

We will illustrate the controlled selection process by an example. The controlled selection arrays used in production in NCS have $M = 152$, $N = 20$ or $N = 23$, and $\gamma = 4$; to keep the illustrative example manageable in size we take $M = 2, N = 3$, and $\gamma = 3$.

There are a number of ways to set up a controlled selection problem. The approach used in (4.1)-(4.6) clearly insures that $\mathbf{S}'$ is an additive array with integer cell values. In Figure 1 the first two arrays presented are the original **S** for the illustrative example and $\mathbf{S}'$ calculated using (4.1)-(4.6). However, there is one drawback to calculating $\mathbf{S}'$ with this approach, which will be addressed at the end of the section.

Other approaches for modifying **S** may not work at all**.** For example, one approach would be to simply round the allocation of each cell in **S**. This generally will not work because the array obtained from rounding **S** will typically not be additive.

We obtain a solution to the controlled selection problem $\mathbf{S}'$ by iteratively constructing a sequence of arrays $\mathbf{A}'_k, k = 1,..., l,$ of dimensions $(M+2)\times(N+2)$ with the marginals of $\mathbf{A}'_k$ an integer multiple of a base b_k, with $\mathbf{N}'_k$ a controlled rounding of $\mathbf{A}'_k$ to the base b_k, and with p_k the probability of selection of $\mathbf{N}'_k$. The set of controlled roundings and associated probabilities satisfy (2.2) without rounding error. The only rounding error occurs in the conversion from **S** to $\mathbf{S}'$.

We begin by letting $\mathbf{A}'_1 = (a'_{ij1}) = \mathbf{S}'$, and $b_1 = 10^{\gamma}$. Then we obtain a controlled rounding $\mathbf{N}'_1 = (n'_{ij1})$ of $\mathbf{A}'_1 = (a'_{ij1})$ to the base b_1. Next we divide each cell in $\mathbf{N}'_1$ by b_1 and round to the nearest integer to obtain $\mathbf{N}_1$. Since $\mathbf{N}'_1$ is an integer multiple of b_1, there should be no rounding error in obtaining $\mathbf{N}_1$ beyond the rounding error in obtaining $\mathbf{S}'$.

Having obtained $\mathbf{A}_1, \mathbf{N}'_1, \mathbf{N}_1$, and b_1, we proceed to explain how for $k > 1$ we obtain by recursion $b_k, p_{k-1}, \mathbf{A}'_k, \mathbf{N}'_k, \mathbf{N}_k$. We let

$$b_k = \max\{\left|n'_{ij(k-1)} - a'_{ij(k-1)}\right| i = 1,..., M+1,\ j = 1,..., N+1\} \tag{4.7}$$

$$p_{k-1} = (b_{k-1} - b_k)/b_1 \tag{4.8}$$

$$a'_{ijk} = (n'_{ij(k-1)} / b_{(k-1)})b_k + a'_{ij(k-1)} - n'_{ij(k-1)} \tag{4.9}$$

$$\mathbf{N}'_k \text{ be a controlled rounding of } \mathbf{A}'_k \text{ to the base } b_k \tag{4.10}$$

$\mathbf{N}_k = (n_{ijk})$ be the array defined by $n_{ijk} = n'_{ijk} / b_k$ with the quotient rounded to the nearest integer (4.11)

Eventually we reach a k for which $b_k = 0$. Then $\ell = k - 1$ and $p_\ell = b_\ell / b_1$, with the tabular arrays $\mathbf{N}_1 = (n_{ij1}),...,\mathbf{N}_\ell = (n_{ij\ell})$ and associated probabilities $p_1,..., p_\ell$ constituting a solution the controlled selection problem $\mathbf{S}'/10^\gamma$.

For the illustrative example, $l = 7$, b_1 - b_8 are, respectively, 1000, 542, 434, 351, 141, 18, 2, and 0 by (4.7) and p_1 - p_7 are, respectively, 0.458, 0.108, 0.083, 0.210. 0.123, 0.016, 0.002. $\mathbf{A}'_k, \mathbf{N}'_k, k = 1,..., 7,$ are given in Figure 1. $\mathbf{N}_1,...,\mathbf{N}_7$ are not presented in the figure but are obvious by (4.11). Note that for this procedure just described, $\mathbf{S}'$ and $\mathbf{A}'_k, k = 1,...l,$ are completely additive integer-valued tabular arrays, which is the key to insuring that the necessary controlled roundings can be obtained. There is rounding error in the construction of $\mathbf{S}'$ from **S**, but it does not destroy any necessary additivity, which could lead to difficulties in performing the controlled roundings. The rounding error in the construction of $\mathbf{S}'$ is at most 1 for any internal cell, which is equivalent to a rounding error of $10^{-\gamma}$ in the original **S.**

The one drawback with the approach using (4.1)-(4.6) is that it does not guarantee that if the grand total for the original controlled selection array **S** was exactly an integer value, that each of the controlled roundings associated with $\mathbf{S}'$ will lead to that grand total. This may or may not be a concern. In the example, the original expected value of the grand total for **S** is 17 but the grand total for $\mathbf{S}'/1000$ is $\sum_{i=1}^{2}\sum_{j=1}^{3} s'_{ij}/1000 = 16.998$ and the grand total corresponding to N_7 is $\sum_{i=1}^{2}\sum_{j=1}^{3} n_{ij7} = 16$, while the grand total corresponding to all the other $\mathbf{N}_k$ is 17.

If this rounding error in the grand total is a concern, it can be avoided as follows. First make the following modification in the construction of $\mathbf{S}'$. In (4.1)-(4.6) replace γ wherever it occurs with $\gamma + \delta$, where δ is the smallest positive integer for which 10^δ is greater than the number of internal cells in **S**. Thus $\delta = 1$ in the illustrative example since there are 6 internal cells in **S**. $\mathbf{S}'$ is presented in Figure 2. Then construct a new tabular array $\mathbf{S}''$ from $\mathbf{S}'$ by first performing a controlled rounding of $\mathbf{S}'$ to the base 10^δ with the additional requirement that $s'_{(M+1)(N+1)}$ be rounded up, not down, in this controlled rounding; and then that each cell in the controlled rounding be divided by 10^δ and rounded to the nearest integer to obtain $\mathbf{S}''$. For the illustrative example $\mathbf{S}''$ is as given in Figure 2. We then let $\mathbf{A}'_1 = \mathbf{S}''_1$ and proceed using (4.7)-(4.11) as we have done previously. The additional requirement that $s'_{(M+1)(N+1)}$ is always rounded up in the controlled rounding can always be satisfied, as explained in Cox and Ernst (1982). In particular, for the illustrative example, $s''_{4,3} = 1000,$ from which it follows that

$$n_{4,3,k} = 1000 \text{ and } \sum_{i=1}^{2}\sum_{j=1}^{3} n_{ijk} = 17 \text{ for all } k$$

Another issue we have is the situation when the controlled roundings have to be selected in a coordinated fashion for two controlled selection problems. In particular, this problem arises when we have separate controlled selection problems for ECI and NCS Wage Only, and we wish to minimize the number of sampling cells for which the sum of the two rounded allocations differs by more than 1 from the expected number of total NCS wage units in the sampling cell. (This can occur if both surveys are rounded in the same direction.) In that case, the construction of $\mathbf{S}'$ is done independently for each controlled selection problem and we set $b_1 = 10^\gamma$. The recursive computation of $b_k, p_{k-1}, \mathbf{A}'_k, \mathbf{N}'_k, \mathbf{N}_k$ for $k > 1$ is done separately and independently for the two surveys using (4.7)-(4.11) with the following exceptions:

After obtaining the controlled rounding $\mathbf{N}'_k$ for ECI, the corresponding controlled rounding for NCS Wage Only is obtained by using an objective function which minimizes the number of cells for which the sum of the roundings for the two surveys differs from the expected value by more than 1. (4.12)

After b_k is computed separately for each survey, the minimum of these two b_k's is taken as the b_k to use in (4.7)-(4.11) for both surveys. (4.13)

See Izsak et al. (2005) for more information on the coordinated selection of the controlled roundings for the two surveys.

5. Base Sample Weights

The procedure for obtaining sample weights is typically more complex when using controlled selection, where the allocation to each sampling cell is not fixed, than for sampling problems where the sample cell allocation is fixed.

Consider a population of N units with weights w_i, values $y_i, i = 1,...,N$, population total $Y = \sum_{i=1}^{N} y_i$ and estimator of total $\hat{Y} = \sum_{i=1}^{N} w_i y_i$. A sufficient condition for this set of weights to result in unbiased estimates of totals is for $E(w_i) = 1$ for each unit (Ernst 1989), since if this condition is met we have $E(\hat{Y}) = Y$. The simplest case for which this condition would be met occurs when the probability of selection p_i of unit i can be calculated for each unit. In this case if we let $w_i = 1/p_i$ when unit i is in sample and $w_i = 0$ otherwise, the set of weights satisfies $E(w_i) = 1$ for all i.

The calculation of p_i is typically easier to do when the allocation of the number of units to each sampling cell is fixed than when it is not. However, there are many situations where this allocation is not fixed. In particular, when controlled selection is used, there are generally two possible allocations for each unit i, which are consecutive integers that we denote by j_i and $j_i + 1$. In this case, there are two possible selection probabilities for the sampling cell containing unit i: p_{ij_i} and $p_{i(j_i+1)}$, corresponding to the allocations of j_i units and $j_i + 1$ units, respectively. Then, provided $j_i \neq 0$, the corresponding weight is $w_i = 1/p_{ij_i}$ if unit i is in sample and the allocation i to the cell containing unit i is j_i units; while $w_i = 1/p_{i(j_i+1)}$ if unit i is in sample and the allocation to the cell containing unit i is $j_i + 1$ units; and $w_i = 0$ if unit i is not in sample. Then under these conditions $E(w_i) = 1$, since $E(w_i) = 1$ conditional on the allocation to unit i's cell being j_i units and also conditional on this allocation being $j_i + 1$ units. (Note that for the NCS Wage sample, we can sometimes have three possible allocations for a cell for reasons discussed at the very end of Section 4 and in more detail in Izsak et al. (2005), but the same weighting idea works in that case too.)

Now, if $j_i = 0$, then unit i is in sample if and only if the allocation to the cell containing unit i is 1 unit and unit i is selected conditional on this allocation. We let r_i be the probability of the former condition being met, while the probability of the latter condition being met is p_{i1}. That is, r_i is the value of the entry in the controlled selection array in the sampling cell containing unit i. Consequently, if we let $w_i = 1/(r_i p_{i1})$ when unit i is selected and $w_i = 0$ otherwise, then $E(w_i) = 1$. Thus r_i is the probability that the allocation to the cell containing unit i is 1 unit and $1/r_i$ is the weighting adjustment to account for the fact that when the allocation to a cell is 0 units, the cell will not contribute to the estimates. r_i is calculated differently for the ECI and the NCS Wage sample. To calculate r_i for the controlled selection problem for ECI, simply set it to the controlled selection value for each internal cell for which the controlled selection value is less than 1; while $r_i = 1$ for all other internal cells. As discussed in Section 3, controlled selection is used in choosing the ECI sample for the government sector, the ECI five panel noncertainty units for the private sector, and the ECI single panel units for the private sector.

For the NCS Wage sample, the calculation of r_i is more complex because the allocation to the Wage sample in a cell containing unit i is greater than 0 if either the index sample or the Wage Only sample has an allocation to that cell that is greater than 0. For each Wage sample to be selected there corresponds an index controlled selection problem and a Wage Only controlled selection problem. For each such pair of controlled selection problems, there corresponds a set of pairs of controlled roundings, one with the index allocation to each cell and the other with the Wage Only allocation, with each pair having an associated probability. For each cell, r_i is the sum of these associated probabilities over all pairs of controlled roundings for which either the index allocation or the Wage Only allocation to the cell is greater than 0. The reason for this is that r_i is the probability that the NCS Wage allocation is positive. Controlled selection is used in choosing the NCS Wage Only sample for the government sector and for the single panel private sector. (Note that in the case when it is possible for a cell allocation to be any of 0, 1, or 2 units, then $w_i = 1/(r_i p_{i2})$ when the allocation is 2 units and unit i is selected.)

6. Minimum Expected Cell Allocations

Now, using the notation of the previous section, if $j_i = 0$ and r_i is very small, then the general tendency would be for w_i to be very large, which typically leads to large variance estimates. To overcome this problem we considered requiring a minimum value for r_i and conducted an empirical investigation comparing five allocation options. Three of these options only differed in the way they calculated the minimum value of r_i. The three values considered in the empirical investigation were $r_i = 0.00$, (that is no minimum), $r_i = 0.01$, and $r_i = 0.05$, which were labeled Options 1, 4 and 3, respectively. Two other options were also considered, Options 2 and 5, neither of which uses minimum values for r_i. In Option 2, unlike any of the other options, minimum

allocations for Pay Agent areas were not considered, while for Option 5, unlike other options, we did not remove the Wage Only sample from nonmetropolitan areas.

We compared the variances of the five options. For national estimates, among the three options that only differed in the value of r_i, Options 3 and 4 had the lowest variances. Option 3 did slightly better than Option 4, likely because of the higher minimum weight adjustment factor, but we preferred Option 4 since we would prefer having minimums to be as small and unobtrusive as possible. However, Option 5 had the lowest national variances among all five options, which we believe is due to the fact that this option is the only option that retains Wage Only units for nonmetropolitan areas. Option 2 had lower national variances than Option 1. This is to be expected since the removal of the Pay Agent area minimums should be expected to lower national variances. However, Option 2 produced higher national variances than Options 3 and 4 since the latter two options use cell minimums.

For the group of Pay Agent areas, variances were fairly similar over the five options, with a slightly higher variance for Option 2, which is to be expected because of the removal of minimum thresholds for Pay Agent areas for this option, and a slightly lower variance for Option 4. For metropolitan areas excluding Pay Agent areas, the average variances were fairly even across the different methods, with Option 5 producing the highest variances and Option 4 producing the lowest. The higher variances for Option 5 for these areas appeared to result from the fact that since it is the only option that does not remove micropolitan and outside CBSAs county clusters from the NCS Wage Only sample, this option has the largest sample for these two types of areas and the smallest sample for metropolitan areas. In micropolitan and outside CBSAs county clusters, Option 5 performed the best for the reason just explained, with no clear pattern for the other four options. The variances for these four options jumped around quite a bit, in part we believe because of the relatively small sample allocated to nonmetropolitan areas for these options.

Note that since each of the options yielded different controlled selection problems, different controlled roundings were used for each of the options. This fact may have resulted in a substantial increase in the variability of the variance estimates for the different options.

It was decided to adopt Option 4. Option 2 was eliminated because it eliminated Pay Agent area minimums, which increased the variances for those areas, without producing the lowest variances for any types of areas. Option 5 was eliminated because of our emphasis on reducing the variances of metropolitan area estimates. Among the other three options, Options 3 and 4 generally produced lower variances than Option 1 for most domains because of the use of minimum real-valued cell allocations. Actually Option 3 produced a slightly lower national variance estimate than Option 4, but the difference was very small and may have been at least partially due to the specific controlled rounding that was selected for each option. In general we thought when in doubt it is better to use a minimum that is small and unobtrusive as possible, while still avoiding very large weight adjustment factors and it was felt that Option 4 best met this requirement.

References

Causey, B. D., Cox, L. H., and Ernst, L. R. (1985). Applications of Transportation Theory to Statistical Problems. Journal of the American Statistical Association, 80, 903-909.

Cox, L. H., and Ernst, L. R., (1982). Controlled Rounding. INFOR, 20, 423-432.

Ernst, L. R. (1989). Weighting Issues for Longitudinal House and Family Estimates. Panel Surveys, 139-159. New York, John Wiley.

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.

Ernst, L. R., Guciardo, C. J, and Izsak, Y. (2004). Evaluation of Unique Aspects of the Sample Design for the National Compensation Survey. American Statistical Association, Section on Survey Research Methods, [CD-ROM], Alexandria, VA: American Statistical Association.

Ernst, L. R., Izsak, Y., Paben, S. P. (2004). Use of Overlap Maximization in the Redesign of the National Compensation Survey. 2004 Proceedings of the American Statistical Association, Section on Survey Research Methods, [CD-ROM], Alexandria, VA: American Statistical Association.

Izsak, Y., Ernst, L. R., Paben, S. P., Ponikowski, C .H. and Tehonica, J. (2003). Redesign of the National Compensation Survey. 2003 Proceedings of the American Statistical Association, Section on Survey Research Methods, [CD-ROM], Alexandria, VA: American Statistical Association.

Izsak Y, Ernst, L. R., McNulty E,. Paben, S. P., Ponikowski, C. H., Springer G., and Tehonica, J. (2005). Update on the Redesign of the National Compensation Survey. 2005 Proceedings of the American Statistical Association, Section on Survey Research Methods, [CD-ROM], Alexandria, VA: American Statistical Association.

Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

$\mathbf{S}=$

1.98424	0.87508	2.59913	5.45845
5.33013	3.91778	2.29364	11.54155
7.31437	4.79286	4.89277	17

$\mathbf{A}'_1 = \mathbf{S}' =$

1984	875	2599	542	6000
5330	3917	2293	460	12000
686	208	108	998	2000
8000	5000	5000	2000	20000

$\mathbf{N}'_1 =$

2000	1000	3000	0	6000
5000	4000	2000	1000	12000
1000	0	0	1000	2000
8000	5000	5000	2000	20000

$\mathbf{A}'_2 =$

1068	417	1225	542	3252
3040	2085	1377	2	6504
228	208	108	540	1084
4336	2710	2710	1084	10840

$\mathbf{N}'_2 =$

1084	542	1084	542	3252
3252	2168	1084	0	6504
0	0	542	542	1084
4336	2710	2710	1084	10840

$\mathbf{A}'_3 =$

852	309	1009	434	2604
2392	1653	1161	2	5208
228	208	0	432	868
3472	2170	2170	868	8680

$\mathbf{N}'_3 =$

868	434	868	434	2604
2604	1302	1302	0	5208
0	434	0	434	868
3472	2170	2170	868	8680

$\mathbf{A}'_4 =$

686	226	843	351	2106
1894	1404	912	2	4212
228	125	0	349	702
2808	1755	1755	702	7020

$\mathbf{N}'_4 =$

702	351	702	351	2106
1755	1404	1053	0	4212
351	0	0	351	702
2808	1755	1755	702	7020

$\mathbf{A}'_5 =$

266	16	423	141	846
844	564	282	2	1692
18	125	0	139	282
1128	705	705	282	2820

$\mathbf{N}'_5 =$

282	0	423	141	846
846	564	282	0	1692
0	141	0	141	282
1128	705	705	282	2820

$\mathbf{A}'_6 =$

20	16	54	18	108
106	72	36	2	216
18	2	0	16	36
144	90	90	36	360

$\mathbf{N}'_6 =$

18	18	54	18	108
108	72	36	0	216
18	0	0	18	36
144	90	90	36	360

$\mathbf{A}'_7 = \mathbf{N}'_7 =$

4	0	6	2	12
10	8	4	2	24
2	2	0	0	4
16	10	10	4	40

Figure 1. Tabular Arrays for Illustrative Example

$\mathbf{S}' =$

19842	8750	25991	5417	60000
53301	39177	22936	4586	120000
6857	2073	1073	9997	20000
80000	50000	50000	20000	200000

$\mathbf{A}'_1 = \mathbf{S}'' =$

1984	875	2599	542	6000
5330	3918	2294	458	12000
686	207	107	1000	2000
8000	5000	5000	2000	20000

Figure 2. Initial Tabular Arrays for Modified Illustrative Example