UPDATE ON THE REDESIGN OF THE NATIONAL COMPENSATION SURVEY August 2005

Yoel Izsak, Lawrence R. Ernst, Erin McNulty, Steven P. Paben, Chester H. Ponikowski, Glenn Springer, Jason Tehonica
Izsak.Yoel@bls.gov, Ernst.Lawrence@bls.gov, Mcnulty.Erin@bls.gov, Paben.Steven@bls.gov, Ponikowski.Chester@bls.gov, Springer.Glenn@bls.gov, Tehonica.Jason@bls.gov
Bureau of Labor Statistics, 2 Massachusetts Ave. N.E., Room 3160, Washington, DC 20212-0001

KEY WORDS: sample redesign, between area variances, allocation of sample, controlled selection

1. Introduction

One of the key products produced by the National Compensation Survey (NCS), which is conducted by the Bureau of Labor Statistics, are locality wage surveys. These wage estimates are produced for metropolitan areas and non-metropolitan areas as defined by the Office of Management and Budget (OMB) in 1994. The NCS surveys two types of metropolitan areas: Metropolitan Statistical Areas (MSAs) and Consolidated Metropolitan Statistical Areas (CMSAs). Non-metropolitan areas are areas that are not part of an MSA or CMSA and are individual counties. In June 2003 OMB released a new set of area definitions. The new area definitions define a set of Core Based Statistical Areas (CBSA) and designate the remaining geographical units as outside CBSA counties. The CBSA areas are divided into Metropolitan Statistical Areas and Micropolitan Statistical Areas. The NCS sample needs to be redesigned to incorporate these new area definitions.

Section 2 of this paper provides a brief summary of the current NCS sample design and an explanation of the old and new OMB area definitions. In Section 3, we present decisions on the sample redesign issues discussed in Izsak et al. (2003), which dealt with our primary sampling units, geographic areas. Section 4 discusses some of the problems in the second stage of sampling, our establishment sample. We will present a solution to these problems, using a controlled selection process to generate our establishment allocations. Section 5 details our establishment allocation process in general, and Section 6 discusses how the actual inputs were determined related to that process.

2. NCS Sample Design

Three of the Bureau of Labor Statistics compensation survey programs, the Employment Cost Index (ECI), the Employee Benefits Survey (EBS), and locality wage surveys, were integrated, creating one comprehensive National Compensation Survey (NCS) program. As a result of this integration, ECI and what was formerly known as EBS share a sample which is a subsample of the NCS Wage sample. The ECI publishes national indexes which track quarterly and annual changes in employers' labor costs and also cost level information, previously annually but now quarterly, on the cost per hour worked of each component of compensation. Annual incidence and detailed provisions of selected employee benefit plans are published in the survey that was formerly known as the EBS. The locality wage surveys program publishes locality and national occupational wage data.

The integrated NCS sample consists of five rotating replacement sample panels. Each of the five sample panels will be in sample for five years before being replaced by a new panel selected annually from the most current frame. The NCS sample 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 probability sample of areas; the second stage is a probability sample of establishments within sampled areas; and the third stage is a probability sample of occupations within sampled areas and establishments.

Currently the NCS sample consists of 152 areas based on OMB's 1994 area definitions. Of the 152 areas, 34 areas were selected with certainty. Three out of the 34 certainty areas would not have been certainty based on total employment, but were added to meet the needs of the President's Pay Agent, a primary customer, because of their large federal employment. (The President's Pay Agent consists of the Secretary of Labor and the Directors of the Office of Management and Budget and the Office of Personnel Management. The Pay Agent makes recommendations for locality pay rates for federal workers.) These 152 areas are comprised of (1) MSAs, areas with a central city of 50,000 or more people and a total area population of at least 100,000, (2) CMSAs, large integrated areas of 1 million or more people consisting of two or more Primary Metropolitan Areas, and (3) non-metropolitan areas, areas that are not part of an MSA or CMSA

3. Area Redesign Issues

Izsak et al. (2003), presented a number of research topics that were being studied in conjunction with the redesign of the area sample for the NCS. The findings presented in that paper were used in the selection process of the new area sample. In this section, we present final decisions for the topics in the 2003 paper.

3.1 Number of sample areas

The 2003 paper studied the optimal number of noncertainty areas to be selected, assuming the number of certainty areas would stay fixed based on the current sample. We decided that the optimal number of non-certainty areas would be determined by calculating the number of establishments and areas that minimized variance for fixed cost. If the proportion of variance due to sampling of PSUs (between PSU variance); is large or the relative costs for sampling an establishment compared

to sampling a PSU are relatively high, then an increase in the number of PSUs is desirable.

The new area sample for the NCS, selected in 2004, has a total of 152 areas. Of the 152 areas, 57 areas were selected with certainty and 95 areas are noncertainty. The total of 152 was based on the current number of NCS wage sample areas. This decision to use 152 areas was based on the priorities and objectives of the NCS, which are to meet current reliability of ECI estimates first, followed by other types of estimates.

3.2 Stratification of the micropolitan areas

Another issue studied in the 2003 paper was the impact on between PSU variances of stratifying the micropolitan areas in different ways. The micropolitan areas could be stratified with the metropolitan areas, with the outside CBSA county clusters, or separately. Variance estimates were calculated for the three scenarios and the results showed minimal differences. Therefore, the final decision was based on what would provide the most flexibility to the NCS for publishing data, which was to stratify the micropolitan areas separately. This gives the NCS the ability to publish separate metropolitan and micropolitan estimates, separate CBSA and non-CBSA estimates, or separate metropolitan and non-metropolitan estimates if so desired.

3.3 Alternative Stratification Variables

A third issue presented in the 2003 paper was the choice of a stratification sorting variable. The old area sample sorted areas for stratification by census division, type of area, and mean wage based on total civilian workers in a PSU. In the new area sample, census division and type of area would be used again, but we wanted to test a few different variables in place of mean wage. Between PSU variances were calculated using a number of different sorting variables including mean wage, weighted sum of industry wages, employment, the proportion of employment in goods-producing industries out of overall employment, and a random sort. Mean wages in a PSU produced the lowest between PSU relative standard errors out of the variables tested. This both affirmed the choice of mean wages in the selection of the old area sample, and served as the basis for choosing mean wages as the sorting variable for the selection of the new area sample.

3.4 Single county vs. multi-county clusters for outside CBSA areas

Another issue studied in the 2003 paper was the possibility of clustering outside-CBSA counties into multi-county clusters. We compared the between-PSU variance over micropolitan areas and outside CBSA clusters (using varying numbers of sample areas for these two area groups) that resulted from the use of single county clusters, and multi-county clusters of varying sizes. We found that clusters with larger employment resulted in lower between-PSU variances. However, geographically-large clusters increase the costs of data collection. In light of these two facts, we decided on clusters containing about 10,000 workers, which are large enough to meet our need to minimize variance, but not so large as to create unwieldy travel costs.

To form the multi-county outside-CBSA clusters of size 10,000 that we decided to use in our sample, we first calculated the total employment and average wage for each outside-CBSA county, over all private and government sector establishments.

Using maps to identify the outside-CBSA counties, along with the county employment and average wage figures, we formed preliminary county clusters using five guidelines:

(1) *County clusters must contain only outside-CBSA counties*.

(2) *Each county cluster must lie within only one Census Division*. The NCS publishes estimates by Census Division, so clusters cannot span multiple Census Divisions.

(3) *Each county cluster must be formed from contiguous counties*. To minimize data-collection travel costs, counties within a cluster should be adjacent.

(4) *Each county cluster should have 10,000 or more workers*. Some outside-CBSA counties with an employment less than 10,000 were surrounded by CBSAs. Isolated single counties with employment between 9,000 and 10,000 were designated as meeting the clustering guidelines, but the preliminary clustering process also left 35 isolated counties with employment less than 9,000 as single-county PSUs.

(5) *Each cluster containing more than one county should be heterogeneous with respect to wage levels*. Heterogeneous PSUs are often preferred in cluster sampling, and in our 2003 paper, we showed that they are acceptable with respect to between-PSU variances. When several clusters were possible in a particular area, we chose county combinations that maximized wage heterogeneity. Heterogeneity within a county cluster was calculated using the following formula:

$$h = \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2$$

where:

n = the number of counties within the cluster

X_i = the mean wage of county i

$\bar{X}$ = the cluster mean wage, calculated by weighting the county means by county employment

With further examination, we realize that a more accurate calculation of heterogeneity might be found by dividing the above formula by n.

Clusters were allowed to cross state borders since the NCS does not publish by state, but we gave first consideration to single-state clusters. When forming the preliminary clusters, we did not consider natural barriers, roads, bridges, commuting patterns, or the nature of the counties' economies, due to resource constraints.

After some revisions, based on input from field economists, the 1,359 outside-CBSA counties were divided among 436 county clusters; 43 multi-county clusters and 22 isolated single-county clusters had employment less than 9,000.

3.5 Maximization of overlap of non-certainty metropolitan areas

In our 2003 paper, we compared the results of several overlap maximization techniques to the results found using no overlap maximization technique. Based on the advantage gained by using overlap maximization, it was decided to use an overlap procedure to select the new sample non-certainty metropolitan areas.

Three different overlap maximization procedures were explored in detail in Ernst, Izsak, and Paben (2004). Here, we will simply summarize the reasons for our decision as to which procedure to use.

One of the overlap procedures considered is the procedure of Causey, Cox, and Ernst (CCE) (1985), which has the key advantage that it yields the true maximum overlap. CCE obtains the optimal overlap by formulating the overlap problem as a transportation problem, a special form of a linear programming problem. Despite this advantage, there are some disadvantages to this procedure that have generally kept it from being used in production, particularly the fact that CCE commonly results in transportation problems that are too large to solve operationally. As a result, we also considered two other procedures, those of Perkins (1970) and Ohlsson (1996), neither of which are difficult to implement operationally. For our particular NCS application, however, we found that the size of the transportation problems in CCE were quite manageable operationally. We selected our new sample PSUs using the CCE procedure, because of the substantially larger expected overlap that it yielded in comparison with the other two procedures that were considered.

3.6 Selecting areas as certainty

We designate as certainty any area with employment greater than 80% of a sampling interval calculated by dividing total employment of all areas in the frame by 152 sample areas. We perform this process iteratively: after all certainty areas are assigned in the first iteration, these areas and their amount of employment are removed from the frame, and the sampling interval is recalculated. We repeat this process until no new certainty areas are found. This resulted in 57 certainty areas in the new area sample.

We use 80% of the sampling interval because areas that are large, but not quite as large as the full sampling interval, can affect the formation of strata and PSU selection for the non-certainty areas. The exact percentage of the sampling interval is set at 80% to preserve continuity with the previous methodology, which also used 80% thresholds.

3.7 Allocating the non-certainty area samples

Our last necessary decision before selecting our non-certainty sample areas was to decide on the method of allocating the number of non-certainty sample areas across the three area types: non-certainty metropolitan areas; micropolitan areas; and outside CBSA county clusters. We tested three different methods of allocating the non-certainty areas across the 27 sampling cells of area type x census division:

1. We looked at an allocation proportional to the total employment within each cell, giving each area group equal opportunity for sample, based on the group's employment.
2. We tried to find an optimal allocation based on minimizing the between-PSU variances. In this method, we set stratum sizes for micropolitan areas and outside CBSA clusters to be smaller than the non-certainty metropolitan strata sizes by a fixed ratio. We tested a range of ratios of non-metropolitan areas to metropolitan areas (the optimal ratio was found to be .45), and also, within the non-metropolitan areas, we tested a range of allocations between the micropolitans and outside CBSA clusters, in order to find the allocation that minimized between-PSU variance.
3. We also considered a compromise between the 45% ratio in method 2 and the 100% ratio in method 1, trying a ratio of .75 between the metropolitan and non-metropolitan areas.

We decided to use method 1 to allocate our non-certainty area sample. Method 2 did produce lower between-PSU variances, but the difference in variance in the three methods was minimal. Therefore, method 1 was chosen because it is straightforward and easy to implement, as well as following the historical preference of NCS for allocations proportional to size.

4. The use of controlled selection to find the NCS Establishment Allocations

4.1 Motivation for the use of controlled selection in NCS

In the NCS wage sample and the ECI sample (a subsample of NCS Wage), our establishments are selected from cells defined by area PSUs × industry sampling strata. There are the same 152 sample PSUs and industry strata for both samples. There are 23 industry sampling strata for the private sector and 20 for the government sector, resulting in 3648 sampling cells for the private sector sample and 3,040 for the government sector. In the government sector, the new wage sample will consist of 4,400 establishments and the index sample will select 2,020 establishments. In the private sector, the single panel sample size will consist of approximately 6150 establishments in the wage sample and approximately 2940 establishments in the index sample. Consequently, if we sampled independently in each

sampling cell with a minimum of one unit in each nonempty cell, then the number of sample units required in order to meet the minimums would be, provided there are few empty cells, more than 1/2 of the total number of sample units for both single panel government and private sector wage samples and more than the total number of sample units for both of the corresponding ECI samples. Thus, sampling independently in each cell results in an inefficient sample for NCS wage, since there is only a small amount of sample remaining to allocate to the large cells after meeting the minimum allocation for each cell. This would also result in an impossible design for the ECI sample, where there is not enough government sample to meet the minimum of 1 in every cell.

One approach to alleviating this problem is to collapse the PSUs into clusters and then treat each cluster as a single PSU for allocation purposes. This is what was done for the previous set of sample panels. In our previous design, there were 54 clusters for the wage sample: 52 clusters consisting of a single actual PSU, one cluster consisting of three actual PSUs from Alaska and Hawaii, and one cluster consisting of 99 relatively small PSUs from the other 48 States. However, even with only 54 clusters, there were still some problems with the allocation, due to the requirement of a minimum of 1 in each cell. In particular, for a number of sampling industries the allocation to the cluster of 99 PSUs was very small, in some cases only a single establishment. Since a very small allocation to cells in this cluster could result in an undesirable increase in the sampling variance, the allocations to these industries were manually increased, which increased the total sample size by approximately 200 establishments.

To avoid a manual adjustment to the allocations and such a large deviation between the target and the actual sample size, we decided to use a different approach based on the two-dimensional controlled selection procedure of Causey, Cox, and Ernst (1985), which guarantees that:

> 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. (1)
>
> The expected number of sample units in each of the domains listed in (1) over all possible samples is the desired number. (2)

The "desired number" for each of these domains or cells is found using a proportional to size allocation, where the sum of the area-weighted frame employments over all units in the domain divided by the same sum over all units in all sampling cells, with this quotient then multiplied by the total sample size, that is a proportional to size allocation. Details on precisely how the cell allocations are obtained are presented in Section 5.

A two-dimensional controlled selection problem in this context is a two-dimensional additive array $\mathbf{S} = (s_{ij})$ in which an internal cell value is the expected number of sample units in the corresponding sample area × industry stratum cell; a row marginal and a column marginal is the expected number of sample units in the corresponding area and corresponding industry, respectively; and the grand total is the total sample size. A solution to a controlled selection problem is a set of integer valued two-dimensional additive arrays $\mathbf{N}_1 = (n_{ij1}),...,\mathbf{N}_\ell = (n_{ij\ell})$ of the same dimensions as **S**, and associated probabilities $p_1,...,p_\ell$ such that for each cell *ij* in each array $\mathbf{N}_k$

$$\left| n_{ijk} - s_{ij} \right| < 1 \tag{3}$$

and for each *ij*

$$\sum_{k=1}^{\ell} p_k n_{ijk} = s_{ij} \tag{4}$$

One of the arrays $\mathbf{N}_k$, which are known as controlled roundings of **S** with each obtained by solving a transportation problem, is chosen with the associated probabilities and the cell values of this array determine the sample allocation to each cell. Note that (1) is satisfied by (3) and (2) is satisfied by (4). The solution to the controlled selection problem involves the solution of a sequence of transportation problems as described in Causey, Cox, and Ernst (1985).

Note that the controlled selection approach avoids the problems associated with cell minimums as follows. Suppose, for example, the expected number of sample establishments in a cell is .3. With a cell minimum of 1, the cell would always be allocated 1 sample unit. With the controlled selection approach, however, the cell would be allocated 1 unit with probability .3 and 0 units with probability .7. Consequently, with controlled selection there would not be an over allocation to very small cells. Also note that when a cell size is fixed, the fixed size must be at least 1, since otherwise units in the cell would have no probability of selection. However, when the cell size is variable, as it generally is with controlled selection, it is acceptable for the cell size to be 0 with a positive probability as long as that probability is not 1.

Since we have two samples, the ECI and the NCS wage, we must solve two different controlled selection problems. We investigated two different routes, described in Sections 4.2 and 4.3 in which we can formulate and solve these controlled selection problems in our allocation process, each with its own positive and negative aspects.

4.2 The use of controlled selection on the ECI sample and NCS total wage sample allocations

The first possibility is to perform a controlled selection on a real valued array of ECI allocations and a real-valued array of total NCS wage (ECI plus wage-only) allocations, with the total NCS wage allocation in each cell at least as much as the ECI allocation The main drawback to this route is that it is possible for the controlled selection to lead to more sample units than the target sample size, when it is rounding the allocations. This can be caused if a rounding is selected where the integer total NCS wage allocation is smaller than the integer ECI allocation in one or more cells.

In any cell where this happens, we would need to increase the NCS wage sample size to equal the ECI sample size, which would create extra sample units. We are unable to control for this in the controlled selection, as there has not yet been found a functional three-dimensional controlled selection process.

A brief example of this problem: Suppose the expected value for ECI sample size in a cell is 2.4, and the expected value for total NCS wage sample size in that cell is 2.6. It is possible for the ECI sample size to be rounded up to 3, while the total NCS wage sample size is rounded down to 2. After the controlled selection process is complete, we would then need to raise the total wage allocation in this sampling cell to 3, creating an extra sample unit. Conceptually, the only way of avoiding this problem would be to add additional constraints that would keep the ECI sample size from being rounded up and the NCS wage sample size to be rounded down simultaneously in cells where the integer parts of both allocations are the same. However such additional constraints would result in a three-dimensional controlled selection problem and it was shown in Causey, Cox, and Ernst (1985) that three-dimensional controlled selection problems are generally not solvable.

It is possible, however, to minimize the number of sampling cells that would have this problem, by modifying the objective function of the controlled selection process. We first labeled all sampling cells where the NCS wage expected value and ECI expected value shared the same whole number portion "problem cells". That is, these cells could present a problem if the ECI value was rounded up and the NCS wage value was rounded down. It is not a problem if this happens in cells where the NCS Wage and ECI expected values have different whole number portions. We set up the objective function to minimize the number of "problem cells" where the ECI expected value would be rounded up. Once a controlled rounding for the ECI portion was complete, we took note of the "problem cells" where the ECI expected value was rounded up, and set up the objective function for the NCS Wage portion up to minimize the number of these cells where the NCS Wage expected value would be rounded down.

We evaluated this approach, using data from the government sector, and there was an expected sample increase of 24, with a possible maximum sample increase of 180 on a target sample size of 4400. This increase was deemed too high,

4.3 The use of controlled selection on the ECI sample and wage-only sample allocations

Another approach to implementing controlled selection in our allocation process would be to perform controlled selection on an array of real valued ECI allocations and an array of real valued wage-only (NCS wage minus ECI) allocations. This will not create an increase in the total sample, as the method in 4.2 would. However, there are other possible drawbacks to this method. It is possible for the real-valued allocations to be rounded up in both the ECI and wage-only samples in the same cell, leading to a total wage allocation for the cell that in some cases is between 1 and 2 units larger than the expected value in that cell. In some cases, this could result in a wage allocation for the cell that is larger than the number of frame establishments in the cell. Similarly, the total wage allocation for a cell could be between 1 and 2 units smaller than expected value, but this could not lead to any issues with respect to the number of frame establishments.

An example of this problem: If the expected value for the ECI allocation in a cell is 0.3, and the expected value for the wage-only allocation in that cell is 0.1, then the rounded total wage sample size for the cell should be either 0 or 1. However, it is possible for both the ECI portion and the wage-only portion to be rounded up to 1, which would result in a total wage allocation of 2. Since other cells are rounded down to compensate for this, it is not a serious problem, unless the number of frame establishments in this cell is 1.

In order to control this problem, we modified the objective functions for the controlled roundings in a similar fashion to our process in Section 4.2. In this case, we also identified "problem cells" in the same way as in the previous section, and we want to prevent any problem cell from being rounded up in both the ECI and wage-only arrays or from being rounded down in both arrays. To minimize the number of problem cells where either of these situations occurs, we allowed the controlled rounding to be performed on the ECI array with no modifications. Then, we examined what happened to the ECI expected values in the problem cells. If the ECI expected value in a problem cell was rounded down, we modified the objective function for the wage-only controlled rounding to minimize the probability that the wage-only expected value would also be rounded down. We made a similar modification going the other way, if the ECI expected value in a problem cell was rounded up.

We ran a trial using this method on government data, and found that the expected number of cells where the allocation would be larger than the frame is 4.6, with the most possible such cells being 9. It was deemed acceptable to reduce the sample by a maximum of 9 sample units, if this occurs, and so it was decided to use the controlled selection in this way for the allocation process.

5. NCS establishment allocation process

Government Sector

In this section we present the steps in our new establishment allocation process for the government sector sample. Our government sample is selected in one pass, whereas the private sample, which will be outlined later, is a bit more complicated, and involves two passes at the selection of a sample.

1. Inputs for the following steps

The inputs needed for the government sector allocation procedure are the total NCS wage sample sizes for the 152 sample areas, and the national ECI allocations for 20 government industry groups. The types of input differ for the two surveys, because the NCS wage survey focuses on locality estimates, but the ECI survey focuses on national estimates, including estimates for the industries.

The ECI government sample is selected from 20 industry strata defined by NAICS (North American Classification System) codes. We calculated the inputs for the national ECI industry allocations proportional to the PSU weighted employment in an industry, out of the total ECI government sample size of 2,020. These inputs are real numbers.

In order to find the inputs for the total NCS wage area sample sizes, we follow the process described in Section 6. In this process, we allocate the total NCS wage government sample of 4,400 units among the 152 sample areas proportional to the PSU weighted employment in each area, with some minimum and maximum sample sizes for certain areas. These allocations are left as real values.

2. Allocation of the ECI sample and NCS wage sample among sampling cells

The sampling cells in the government sector for both NCS Wage and ECI are the intersections between the 20 government industry strata and the 152 sample areas. This results in 3040 government sampling cells.

The next step in our allocation process is to find ECI sample allocations for each of the 3040 sampling cells. These allocations are found by allocating each national ECI industry allocation, as found in step 1, among the 152 sample areas, proportional to the weighted employment in each area within the industry. These allocations will be real valued.

We then find NCS Wage allocations for each area x industry sampling cell. To calculate these, we take the total NCS Wage area sample size found in step 1, and allocate these area sizes among the 20 industry strata, proportional to the weighted employment in each industry within the area. These allocations are also real valued.

3. Adjustments to the NCS wage allocations

Because ECI is a subsample of NCS Wage, we need to ensure that in all sampling cells, the NCS Wage allocation found in step 2 is greater than or equal to the ECI allocation. If any sampling cell is assigned an ECI allocation in step 2 that is greater than the NCS Wage allocation, then we raise the NCS Wage allocation to be equal to the ECI allocation. Once this has been done, we remove this cell from further consideration, and subtract the amount of sample and employment in this cell from the respective total numbers. Then, we repeat step 2, with the remaining sampling cells. This is done iteratively, until no cell has an NCS Wage allocation smaller than its ECI allocation.

A similar problem can occur, in that the ECI allocation and/or NCS Wage allocation in a cell, as found in step 2, could exceed the frame size in that cell. If this is the case, a process similar to the preceding paragraph is followed, where the ECI allocation and/or NCS Wage allocation is lowered to the frame size, and the other sample cells are reallocated, until no cell allocations exceed the frame size in that cell.

4. Finding integer allocations for each sampling cell

In order to find the final government sector sample allocations, the real valued allocations resulting from steps 2 and 3 need to be converted into integers. To convert the real value allocations to integers, a two-dimensional controlled selection, as described in Section 4.1 and 4.3, is used on two arrays of allocations: the real valued array of ECI allocations, and a real valued array of "wage-only" allocations, which is found by subtracting the ECI allocation in a sampling cell from the NCS Wage allocation in a cell. Once these integer allocations are found for the index and wage-only sample numbers, we sum the integer ECI allocation in a cell with the integer wage-only allocation in a cell to arrive at the integer NCS Wage allocation in that cell.

As mentioned in Section 4.3, it is possible for the controlled selection process to assign some cells an integer NCS Wage allocation that is 1 unit larger than the number of frame establishments in that cell. Where this happens, the extra unit will be cut from the sample. As mentioned in Section 4.3, the maximum amount of sample that would need to be cut in this way in the government sector is 9.

Private Sector

This section will detail the allocation process for the private sector. In steps that are similar to the government sector's process, this will be noted, and the difference between the government and private sectors will be discussed.

1. Inputs for the following steps

The private sector has 23 industry strata, as opposed to the government sector's 20. Also, the private sector has a national ECI sample size of 15,980, and a national NCS Wage sample size of 32,890. These national sample sizes were based on the number of establishments that BLS can collect, with respect to cost and work hours. We find the NCS Wage total area allocations in a similar manner to the government sector's process. However, the ECI national industry allocations, which will be used as inputs, are based on historical data and are designed to meet certain variance objectives. These are discussed in further detail in Section 6.

2. Allocation of the ECI sample and NCS wage sample among sampling cells

These allocations are found in an identical manner to step 2 in the government sector allocation process.

3. Adjustments to the NCS wage allocations

This step is also implemented in an identical manner to step 3 in the government sector allocation process.

4. Selecting the NCS Wage and ECI certainty units

The private sector sample selection differs from the government sector's sample selection, in that the private sector sample is selected in two passes through the frame. This is because while the entire government sample is selected at one time and is intended to remain fixed for at least five years, the private sector sample is a rotating panel sample with approximately 1/5 of the entire sample selected each year. However, the private sector certainty units are for the most parts selected to be in sample for five panels and are selected using sampling intervals corresponding to the full sample size. The reason that certainty units are selected this way is explained in Ernst , Guciardo, and Izsak (2004). The first pass principally

selects the five-panel units in the sample, and the second pass principally selects the single-panel units. This step outlines the first pass sample selection, in which we select the five-panel certainty units for both the NCS Wage and ECI samples.

An NCS Wage sampling interval is calculated for each sampling cell by dividing the cell's PSU weighted employment by the NCS Wage allocation to that cell, found in steps 2 and 3. Any establishment in the cell with weighted employment greater than the sampling interval is found to be a NCS Wage five-panel establishment. These wage five-panel establishments are then removed from the sampling frame, their employment is subtracted from the total employment in the cell, and the cell's allocation is reduced by the number of five-panel establishments already found in the cell. A new sampling interval is calculated, and this step is repeated iteratively until no new five-panel establishments are found in a cell.

The ECI five-panel certainty establishments are then found in an identical fashion. However, since the ECI allocation in any cell cannot exceed its NCS Wage allocation, any establishments found to be ECI five-panel certainty must have been found to be NCS Wage five-panel units in the step corresponding to the previous paragraph.

5. ECI five-panel noncertainty sample sizes

This step marks the beginning of the second pass in the selection of NCS Wage and ECI samples. For the second pass, the allocations among cells of single-panel sample sizes for NCS Wage and ECI must be determined, as well as the allocation of five-panel ECI noncertainty units.

For NCS Wage, all five-panel units are certainty units. In ECI, some of the establishments that were selected as NCS Wage five-panel will be selected as ECI noncertainty units, because of the smaller sample size for ECI. If these units were assigned to a single panel in ECI, this would increase respondent burden on these units, since a wage initiation would be required in the first panel, and then a benefits initiation would be required for ECI in the panel the unit was assigned to in ECI. Because of this, any NCS five-panel unit that was selected in ECI would need to be selected for all five-panels in ECI. So, to solve this problem, the second pass ECI allocation is split into two parts: five-panel noncertainty sample sizes and single-panel sample sizes. (Ernst et al. 2002)

So, the possible ECI five-panel noncertainty units include any unit that is a NCS Wage five-panel unit, but was not selected as an ECI five-panel certainty unit in step 4. In order to find the ECI five-panel noncertainty sample allocation to a sampling cell, first, the number of ECI certainty units in the cell is subtracted from the overall ECI sample size in that cell. This results in the total noncertainty sample size for that cell. The employment in all possible five-panel noncertainty units (those units that were selected as five-panel units in NCS Wage but not as certainty in ECI) makes up the "five-panel noncertainty frame employment". The single-panel frame employment in a cell is found by subtracting the frame employment of all units selected as five-panel units in NCS Wage. Then, the total ECI noncertainty sample size is allocated among the group of five-panel non-certainty units and the single panel units, proportional to PSU weighted employment. The five-panel non-certainty allocation and single panel allocations in all cells will be real valued, at this point in the process.

In order to find integer allocations for the ECI five-panel noncertainty units in each sampling cell, we use a controlled selection process on the array of real valued allocations. Since five-panel noncertainty units only exist for ECI, only one controlled selection array is needed in this step instead of the ECI and wage only arrays needed for the government and the private sector single-panel allocations.

6. NCS Wage and ECI single-panel allocations

To obtain the NCS Wage single-panel allocation over all five panels in a sampling cell, the number of NCS Wage five-panel areas in the cell (found in step 4) is subtracted from the total NCS Wage sample allocation found in step 3. This number is then divided by 5, to arrive at the single-panel sample allocation to each of the five panels. These allocations will be real numbers at this point.

The ECI single-panel allocations over all five panels were found in step 5, and these also are divided by 5 to arrive at the single-panel sample allocations to each of the five panels. These are also real-valued allocations at this point.

At this point, it is possible for the ECI single-panel allocation in a cell to exceed the NCS Wage allocation in that cell. If this occurs, the NCS Wage allocation is raised to the value of the ECI allocation, and all other industries in that area are reallocated. This is done in a similar manner to step 3.

A controlled selection is then performed, as in step 4 of the government sector process, on the array of real-valued ECI single-panel allocations and the array of real-valued wage-only allocations. This controlled selection process differs slightly from those used in the previous steps. In this step, we will select using their associated probabilities five independent controlled roundings, one for each of the five-panels, from the set of controlled roundings found in solving the controlled selection problem. We then sum the resulting integer ECI and wage-only allocations in each panel, to obtain the integer NCS Wage allocations in each of the five panels.

Since we process the 5-panel non-certainty ECI allocations and the ECI single-panel allocations in separate controlled selections, it is possible that the sum of the integer 5-panel non-certainty allocation and the integer single-panel allocation in a cell could be greater than the total non-certainty allocation in that cell.

Also, we cannot control for the sum of the 5 single panels, so a similar problem could occur where the 5 single panel allocations do not sum to the total single panel allocation. However, we have this same problem in the current sample. This problem is actually worse in the current process, because if a single-panel allocation is rounded down (or up) in one panel, it is rounded down (or up) in all 5 panels, whereas, if we use our controlled selection process, it is highly unlikely that the sum of the 5

single panels in a sampling cell will deviate from its expected value by 4.

6. Calculation of the inputs to the allocation process

This section discusses the calculation of both the ECI national industry allocations and the NCS wage total area allocations, used as inputs in step 1 of both the private sector and government sector allocation processes.

When allocating the ECI sample, we first allocate the total national ECI sample among the industry strata, due to the fact that the focus in the ECI survey is on national estimates, including estimates for industry and occupational groups. In the private sector, we will use the national ECI industry allocations that were used for our most recent sample, thus preserving the oversampling or undersampling of certain industries that was done previously. One reason for oversampling certain industries in the private sector is that some industries contain a large number of incentive based jobs, which creates a large variance in those industries. This problem does not exist in government, so we do not oversample or undersample any industries in the government sector. For the government ECI sample, we will allocate the national ECI sample size across the 20 government industries proportional to the PSU weighted employment in each industry.

The NCS wage sample begins from the opposite direction, allocating the national wage sample among the sample areas first, due to the fact that the main focus of the NCS Wage survey is on locality estimates. For the private sector NCS wage sample, we examined a few different methods of finding the total area sample allocations:

1. Allocations proportional to PSU weighted employment;
2. Allocations proportional to the square root of PSU weighted employment;
3. Allocations proportional to PSU weighted employment raised to a power of 0.4. (This power of 0.4 was determined to be the optimal exponent through a power regression.)

A decision was made to use allocations that were proportional to PSU weighted employment, as this method produced allocations that were fairly similar to the current total NCS Wage area allocations in the most important areas. However, this method produced some very large allocations for the three largest areas, New York, Los Angeles, and Chicago, and some allocations that were quite small for certain Pay Agent requested areas such as Hartford, Dayton, Richmond, and Huntsville. There are some areas with large federal workforce that are added to the list of certainty areas at the request of the President's Pay Agent. In certain areas such as these three, the private sector workforce is not very large, but it is necessary to assign a relatively large sample allocation to these areas, to meet the needs of the Pay Agent. To correct the discrepancies between the previous sample's allocations and these new allocations, for the largest and smallest Pay Agent requested areas, we decided to specify minimums of 250 units for any Pay Agent areas, and a maximum of 1300 for New York and of 1200 for Los Angeles and Chicago. The minimums were based on past desired minimums for Pay Agent areas, in order to obtain a similar amount of reliability to that in the past. The maximum of 1300 for New York is based on the current sample's New York ECI sample size of 1304. After assigning this maximum to New York, the sample sizes for Los Angeles and Chicago were greatly inflated, so it was necessary to assign maximums to those two areas also. The maximums for Los Angeles and Chicago were set below New York's maximum, but larger than the next largest certainty area.

For the government sector's NCS wage sample, we use a method similar to that of the private sector sample. In the case of the government sector, the minimum total area allocation for any Pay Agent area was assigned as 30, and the maximums assigned were as follows: 152 to New York, 130 to Los Angeles, and 112 to Chicago.

7. 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.

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.

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