REDESIGN OF THE NATIONAL COMPENSATION SURVEY

Yoel Izsak, Lawrence R. Ernst, Steven P. Paben, Chester H. Ponikowski, Jason Tehonica
Izsak.Yoel@bls.gov, Ernst.Lawrence@bls.gov, Paben.Steven@bls.gov, Ponikowski.Chester@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, Sample areas, Between area variances, Overlap maximization, Allocation of sample

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). Individual counties that are not part of an MSA or CMSA are considered non-metropolitan areas. 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 will need to be redesigned to incorporate the 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 discuss some of the sample redesign issues that we are studying, including impact on between area variances of (1) using single outside Core Based Statistical Areas (CBSAs) counties versus multi-county clusters outside of CBSAs as primary sampling units (PSUs), (2) stratifying micropolitan areas with metropolitan areas, or micropolitan areas with counties outside of CBSAs, or these three types of areas separately, and (3) alternative stratification variables. Additional issues discussed are: total number of sample areas and establishments, allocation of the total number of non-certainty areas among metropolitan areas and non-metropolitan areas, and the use of an overlap maximization procedure to select non-certainty metropolitan areas. Section 4 will present the results of the empirical studies corresponding to each of the redesign issues. Finally, Section 5 will list transition and other future research issues.

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. 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. The EBS publishes annual incidence and detailed provisions of selected employee benefit plans. 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 154 areas based on OMB's 1994 area definitions. Of the 154 areas, 36 areas were selected with certainty. These 154 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 (large areas that consist of 250,000 to 999,999 people), and (3) non-metropolitan areas, areas that are not part of an MSA or CMSA. The new OMB area definitions define CBSAs and outside CBSA counties. A CBSA is a geographic entity associated with at least one core of 10,000 or more population, plus adjacent territory that has a high degree of social and economic integration with the core as measured by commuting ties. The CBSA areas are divided into Metropolitan Statistical Areas and Micropolitan Statistical Areas. Metropolitan Statistical Areas, are areas based on urbanized areas of 50,000 or more population, and Micropolitan Statistical Areas, are based on urban clusters of at least 10,000 population but less than 50,000 population.

As part of the redesign process, the NCS program will re-select its sample of PSUs using 2003 CBSA and outside CBSA definitions. This will replace the current set of PSUs defined by MSAs and non-metropolitan counties.

3. Redesign Issues Investigated

We present here short descriptions of the redesign issues that we have investigated to date.

3.1. Impact on between PSU variances of single county versus multi-county outside CBSA areas.

Currently in the NCS program each county that is not part of an MSA is considered a separate PSU. There are some drawbacks to this approach. Some of these counties have very small employment leading to large PSU sampling weights, which can adversely affect variances. Also such

small PSUs do not result in heterogeneous PSUs, which is frequently considered a goal in cluster sampling due to the resulting smaller intra-class correlation coefficients. Finally, one of the goals of the NCS program is to publish wage data for as many of the sample areas as possible. More publications might be possible with multi-county outside CBSA PSUs, but only if we can reduce the number of sample outside CBSA PSUs, which is discussed in Section 3.4 below. The decision on the single county versus multi-county issue will be made primarily on the basis of which approach produces lower between PSU variances (the component of variance arising from the sampling of PSUs), and we consequently compare the between PSU variances for both approaches.

We also investigated the issue of whether heterogeneous or homogeneous multi-county PSUs result in lower between PSU variances. Although, as mentioned above, heterogeneous PSUs are considered preferable in cluster sampling, the general goal in forming strata is to have all the PSUs in each stratum be as similar as possible, and it is not clear whether these two goals work well together.

Throughout this paper the calculation of between PSU variances was done using data directly from our Longitudinal Data Base (LDB) frame. The LDB is the sampling frame for the NCS Program. This has the advantage over using NCS sample data of being able to calculate between PSU variances under PSU definitions other than the current ones and avoids the problems of negative estimates of between PSU variances that can occur when estimating between PSU variance as the difference of estimates of total variance and within PSU variance from sample data. It has the disadvantage that we cannot calculate between PSU variances for occupational estimates since there is no occupational information on the frame.

3.2. Impact on between PSU variances of stratifying micropolitan areas with: metropolitan areas, outside CBSA areas, and separately.

Previously, all non-certainty PSUs were either MSAs or non-metropolitan counties and we stratified these two types of PSUs separately. The new OMB definitions will now include three types of non-certainty areas, metropolitan, micropolitan, and outside CBSAs. Should we stratify the micropolitan areas with metropolitan areas, with the outside CBSA areas, or form three types of strata by stratifying the micropolitan areas separately? We compare the between PSU variances under these three options.

3.3. Alternative stratification variables.

In the current design, the stratification variables used were census division, type of area (metropolitan or non-metropolitan) and mean wages of all civilian workers in a PSU, which can be calculated from our LDB frame data. We investigate the impact on between PSU variances of several alternatives to mean wages as a stratification variable, all of which can also be calculated from the frame.

3.4 Allocation of the total number of non-certainty areas among metropolitan and non-metropolitan areas.

There are two sub-issues here. First, for any fixed total number of non-certainty areas, what is the optimal allocation between these two types of areas in terms of minimizing between PSU variance for national domains? Also, how do the various allocations affect the between PSU and total variances of separate metropolitan and non-metropolitan estimates? The second issue is more important for the non-metropolitan estimates since the employment in non-metropolitan areas is much smaller than for metropolitan areas and the total variance consequently larger.

In the current design there are actually many more sample non-metropolitan PSUs, than there are non-certainty metropolitan PSUs, 73 vs. 45, despite the fact that the employment in metropolitan non-certainty strata is much larger than in non-metropolitan strata. This work will indicate whether such an allocation is desirable. If the number of non-metropolitan sample PSUs can be reduced while the total number of sample establishments in all non-metropolitan sample PSUs is held constant, then the larger average number of sample establishments per sample PSU should allow for the publication of more non-metropolitan areas.

3.5. Total number of sample areas.

This is actually more of an issue of the total number of non-certainty areas, since we anticipate that the total number of certainty areas will not change much from the current design. This is because the establishments in most of the certainty areas are over-sampled for wage estimates relative to the proportion of the employment in these areas in order to insure the reliability of the estimates in these areas. The tradeoff is that this decreases the reliability of national and census division estimates. Consequently, given our budget constraints, we do not believe we can increase the number of certainty areas appreciably. We have been assuming so far in our research that the number of certainty areas and the total number of sample establishments in certainty areas will remain unchanged from the current design, although there may be some increase due to the needs of the President's Pay Agent.

There are two main factors that determine the optimal number of non-certainty sample areas. First is the proportion of total variance that is due to sampling of PSUs (between PSU variance). In general, if this proportion is large, then an increase in the number of PSUs is desirable to reduce this component. This proportion is calculated for national estimates. The other factor is the relative size of the components of the survey costs that are functions of the number of sample PSUs in the design and of the number of sample establishments. The first component includes costs such as travel costs to and from PSUs, while the second component includes costs such as the time spent in directly collecting the data and the travel costs within a PSU. The larger the relative size of the first component, then the smaller the number of sample PSUs and the larger the number of total sample establishments needed to minimize variance for fixed costs. We calculate the optimal number

of sample PSUs and establishments under a very simple cost model for various values of the model parameters.

3.6 Maximization of overlap of non-certainty metropolitan areas.

This is a very different topic. A number of survey programs maximize the overlap of PSUs when reselecting an area sample in order to minimize the extra expenses associated with replacement of sample areas. In addition, in the case of NCS, the more continuing sample areas that are in sample in the new design, the easier it will be to publish for a large number of areas during the transition to the new area design. To estimate the gains in expected overlap from using an overlap maximization procedure, we simulated new stratifications for the non-certainty metropolitan areas and compared the gains in expected overlap using two simple overlap maximization procedures to the expected overlap if the new sample areas are selected independently of the current sample areas, that is, without an overlap procedure. Since the new PSUs corresponding to the current non-metropolitan PSUs may be very different, as described in Sections 3.1 and 3.2, we have not studied the impact of overlap maximization for these PSUs.

4. Results and Conclusions

In Sections 4.1-4.6 we present the results of empirical studies corresponding to the research issues described in Sections 3.1-3.6.

The research issues described in Sections 3.1-3.5 each require the formation of a set or multiple sets of sampling strata corresponding to a universe of PSUs. The algorithm used to form the strata is described in Appendix A. Each stratification depends on a stratification variable, which is fixed as mean wage except in Section 4.3, and also on the number of sample PSUs by groups, such as non-certainty metropolitan PSUs and other non-certainty PSUs.

Next, a between PSU variance is calculated, corresponding to each stratification and each estimate. The methods we used to calculate the between PSU variances are described in Appendix B. All references to variances in Sections 4.1-4.4 are to between PSU variances.

We now proceed to present the results of the empirical studies.

4.1. Impact on between PSU variances of single county versus multi-county outside CBSA areas.

Prior to forming multi-county outside CBSA PSUs to compare against single county PSUs, we decided to test whether heterogeneous or homogeneous multi-county clusters would result in lower between PSU variances.

Non-contiguous heterogeneous and homogeneous multi-county clusters were used as PSUs, and created by algorithms described in Appendix C, separately in each census division.

The between PSU variances that were computed for the homogeneous and heterogeneous multi-county PSUs are included in Table 1. All of the between PSU variance estimates in this paper were computed using frame data from the last quarter of 2001, and are for estimates of quarterly mean wages. All estimates in this section are restricted to workers in the private sector. The RSEs throughout this paper are relative standard errors expressed as percents.

The RSEs in Table 1 and some of the other tables are weighted averages. RSEs for All Workers and major industry divisions were not included because of the length requirements for this paper. To calculate these weighted averages, we first squared the RSEs for All Workers and 6 major industry divisions: Construction, Manufacturing, Transportation, Retail & Wholesale Trade, FIRE, and Service (we omitted Mining, because its employment is very small) to create relative variances. Then, the relative variance for All Workers was multiplied by 6 to give greater weight to the All Workers estimate. This weighted relative variance for All Workers and the relative variances for the 6 industries were then summed, and the resulting sum was divided by 12, to take into account the total weight. Then, the square root of this result was calculated, which gives the weighted mean RSE that we used in the tables. (This final result is not actually a RSE, but we will label it that way.)

The variances in Table 1 were estimated for data consisting only of the non-metropolitan areas, using 73 non-metropolitan strata, the current number. Variances were also computed for other amounts of non-metropolitan strata ranging from 60 to 80, with similar results. The All Workers RSE and the weighted mean RSE are slightly lower for the heterogeneous county clusters, but three of the industry divisions have lower RSEs with homogeneous county clusters. This suggests that there was no clear advantage for either homogeneous or heterogeneous clusters, so we decided to create PSUs in the standard way, and combine counties heterogeneously.

Table 1: Homogeneous vs. Heterogeneous RSEs

Homogeneous multi-county clusters	Heterogeneous multi-county clusters
1.38	1.35

Multi-county PSUs with minimum employment size of 3,000 were then created heterogeneously from the preliminary universe of outside CBSA counties. The PSUs formed were contiguous, with no major physical impediments to travel between counties by automobile. Between PSU variances were then computed for these multi-county PSUs and compared to the variances for single county PSUs. These variances were compared in each census division, and for national data. Table 2 includes weighted mean variance estimates for national data, using 73 non-metropolitan strata. Similar results were obtained for other amounts of non-metropolitan strata ranging from 58 to 88. The multi-county PSUs produced a lower weighted mean RSE than the single county PSUs, and this result was consistent in all industry domains except Retail and Wholesale Trade. These results suggest that it would be advantageous to assign the outside CBSA areas to multi-county PSUs in the new design.

Table 2: Multi-County vs. Single County PSU RSEs

Single County PSUs	Multi-County PSUs
1.58	1.52

Multi-county PSUs with targeted employment larger than 3,000 are also being considered, and the issue of finding an optimal target employment for the multi-county PSUs will be studied in the near future.

4.2. Impact on between PSU variances of stratifying micropolitan areas with: metropolitan areas, outside CBSA areas, and separately.

Between PSU variances were computed for three different scenarios regarding the classification of micropolitan areas:

1. The micropolitan areas grouped with the outside CBSA counties.
2. The micropolitan areas grouped with the non-certainty MSAs.
3. The micropolitan areas as a group of their own.

The group of micropolitan areas that was examined included all micropolitan areas that were not assigned to a certainty area. Scenario 1 resembles the current area design more than the other scenarios. In particular, the intersection of the current non-metropolitan group and the group defined by scenario 1 (using the new area definitions) includes counties comprising 91% of the current non-metropolitan areas and 90% of the group defined by scenario 1, by employment.

Variances were computed for these three scenarios over national data and the census divisions, and some of the results are included in Table 3. These variances were computed using multi-county PSU assignments for the outside CBSA counties.

Table 3 includes weighted average between PSU RSEs for national estimates using the optimal allocation of strata for each scenario. That is, the allocation of strata used minimizes the RSE for each scenario, using the current total of 118 non-certainty strata. The methodology used for obtaining the optimal allocation is described in Section 4.4. Variances were also estimated for other allocations and other numbers of total strata and produced similar results.

Table 3: Various classifications of micropolitan areas

Scenario	# MSA strata	# non-met strata	# micro strata	RSE
1	65	53	0	0.74
2	80	38	0	0.77
3	65	28	25	0.80

Table 3 shows that for the national data, scenario 1 produced slightly lower variances than scenarios 2 and 3, but the differences in the three scenarios' RSEs are minimal. Currently, the non-metropolitan area estimates that we publish are for a set of counties that are closest to scenario 1. Consequently, for this reason and because scenario 1 had a slightly smaller variance than the other two scenarios, we will group the micropolitan areas with outside CBSA counties in the remainder of the paper. No final decision has been made as yet on which scenario will be adopted in the redesign.

4.3. Alternative stratification variables.

To create the current area design, the PSUs were sorted by mean wage before assignment to sampling strata. However, it is possible that sorting by a different variable would create sampling strata with lower between PSU variances. Variances were computed using the following variables, for the purpose of comparison:

1. Mean Wage
2. Weighted sum of Industry Wages
3. Employment
4. The proportion of employment in Goods-Producing industries out of overall employment
5. A random sort

Variable 2 was used to test the incorporation of more domains (the industries) into the creation of strata. Variables 3 and 4 were used to test the influence of employment, which has been used as a stratification variable in some past redesigns, and variable 5 tested the necessity of using any sorting variable at all.

The variance estimates that were computed by using these five stratification variables are included in Table 4. The weighted mean between PSU RSEs in Table 4 were computed for national estimates. The number of strata used for the results in Table 4 was the present allocation of 45 non-certainty MSA strata, and 73 non-metropolitan strata, with the outside CBSA counties grouped into multi-county PSUs. Other amounts of strata produced similar results. The RSE for stratification variable 1, mean wages, was lower than the RSEs of any of the alternative stratification variables. This affirms the use of mean wage as the stratification variable.

Table 4: Alternative Stratification Variable RSEs

Var 1	Var 2	Var 3	Var 4	Var 5
0.87	1.33	1.34	1.22	1.38

4.4 Allocation of the total number of non-certainty areas among metropolitan and non-metropolitan areas.

Between PSU variances were calculated for a range of total non-certainty strata from 98 to 138 in increments of five. For each total number of non-certainty strata, the allocation of non-certainty MSA strata ranged from 35 to 90 in increments of five, with the remaining number of non-certainty strata belongng to the non-metropolitan areas. The optimal allocation for each total number of strata was determined by comparing the between PSU variances for the different allocations within these ranges. Between PSU variances were calculated over all non-certainty areas, with the outside CBSA counties grouped into multi-county PSUs.

Table 5 contains weighted mean between PSU RSEs, as described in Section 4.1. The first row includes the RSE for the current allocation of 45 non-certainty MSA strata and 73 non-metropolitan strata. The remaining rows have RSEs corresponding to the optimal allocations of non-

certainty strata between metropolitan and non-metropolitan PSUs for each total number of strata within the range tested.

Variances decreased as the number of total strata increased. Also, the optimal allocation between the metropolitan and non-metropolitan areas in all numbers of total strata generally indicated that metropolitan areas should make up a higher proportion of the non-certainty sample areas than in the current design.

Table 5: RSEs for various allocations of strata

Total strata	MSA strata	Non-met strata	Weighted Mean RSE
Current Allocation			
118	45	73	0.868
Optimal Allocations at each number of total strata			
98	60	38	0.863
103	60	43	0.826
108	60	48	0.792
113	60	53	0.765
118	65	53	0.739
123	65	58	0.719
128	75	53	0.692
133	75	58	0.672
138	75	63	0.652

4.5 Optimal number of total sample areas and establishments.

We calculated for the nation, the total number of sample areas and establishments in the sample that would minimize total variance for fixed costs.

First, however, we calculated ratios of the between PSU variance to the total variance, for national data for our current design, which are presented in Table 6. Generally, the higher these ratios are, the greater the need for more sample PSUs. The total variances in Table 6 were calculated using the current BRR variance estimation programs for NCS, with NCS sample data obtained from the current design. Because the total variance was calculated using current area definitions, the corresponding between PSU variances were also calculated using the current area definitions, most notably by using single county PSUs, and with the micropolitan areas not defined as such. The variances in the first two columns of Table 6 are relative variances.

Table 6: Ratios of between PSU to total variances

	Total Rel Var	Bet-PSU Rel Var	Ratio of Between to Total Var
Weighted Mean	3.378	0.763	0.226

The ratio of the between PSU variance to the total variance is fairly small. This ratios is approximately what we have been assuming in past sample redesign work, an indication that the current number of sample PSUs may be about right.

However, the data in Table 6 by itself is not enough to determine the optimal allocation of the number of PSUs versus the number of establishments. To accomplish this, total, between PSU, and within PSU variances were used together with a cost model while varying the values of a parameter of the cost model. The variances that were computed are included in Table 7 and Table 8.

We used the following simple cost model:

$$C_T = C_F + (C_A n_A) + (C_E n_E) \quad (1)$$

- where C_T, C_F, C_A, and C_E = total cost, a fixed cost, the cost per area, and the cost per establishment, respectively
- n_A, n_E = number of areas and total establishments

C_F was subtracted from (1), and the result was divided by C_A, yielding:

$$T = C_R n_E + n_A \quad (2)$$

- where $T = \dfrac{C_T - C_F}{C_A}$
- $C_R = \dfrac{C_E}{C_A}$

We used the previously calculated weighted mean of the total relative variance (3.378, from Table 6). The weighted mean, 2.615, of the within PSU relative variances for the current private sector sample size of 37,284, was calculated by subtracting the entry in the second column of Table 6 from the first column. This will be used in step 3 below.

The between PSU variances in Tables 7 and 8 are relative variances for the optimal allocation for the indicated values of n_A (calculated using the allocations in Table 5, but now using single county PSUs, and n_A is 36 more than the corresponding total number of strata in Table 5, since n_A includes the certainty PSUs). The following steps were carried out for various values of C_R, n_A, and n_E:

1. T was calculated from (2), using the current values for n_A (154), and n_E (37,284), and a range of values for C_R (from .005 to .5). (A range was used because we do not have information yet on the actual value of C_R). For each value of C_R, the value for T remained constant in future steps.
2. For a value of C_R, we used a range of values for n_A from 134 to 204 in intervals of 5, and for each value solved (2) for n_E
3. The (original) within PSU relative variance of 2.615 was then adjusted by multiplying by $37{,}284/n_E$. (We assumed that within PSU variance was inversely proportional to the number of establishments)
4. Finally, an adjusted total variance was calculated (for each set of values of C_R, n_A and n_E) by adding the between PSU relative variance for the value of n_A, and the within PSU relative variance that was calculated in step 3 for the value of n_E.

The variances in Table 7 and Table 8 are all weighted mean relative variances. Table 7 gives variances for a C_R value of 0.005, for a range of n_A from 134 to 204, which shows that for this cost ratio, the optimal number of total areas (with respect to total variance) is 139. In Table 8, variances are included for each value of C_R but only for the number of areas that minimizes total variance for that value of C_R. Note that optimal value of n_A increases as C_R increases, since it becomes relatively less expensive to sample more areas rather than more establishments. Also with $C_R \geq .05$, the optimal value of n_A in our range is 204, the maximum value we calculated. If we calculated higher values for n_A, it is likely we would have obtained higher optimal values for n_A across the range of cost ratios.

Each entry in the final column in Tables 7 and 8 is the ratio of the total relative variance calculated in step 4 to the original total relative variance (3.378) that was based on our current number of establishments, metropolitan sample areas, and non-metropolitan sample areas. For each value of C_R it indicates the proportional amount of variance reduction obtained from using an optimal allocation rather than the current allocation. Note that the entries in this column do not vary much, indicating that the optimal variance is not very sensitive to the value of C_R.

In Tables 7 and 8 we use the following abbreviations:

V_B = Between PSU relative variance

V_W =Within PSU relative variance

V_T =Total relative variance

$R_T = \frac{V_T}{3.38}$ (3.38 is total variance for the current design)

Table 7: Number of areas vs. variance

C_R	n_A	n_E	V_B	V_W	V_T	R_T
0.005	134	41284	0.765	2.361	3.126	0.925
0.005	139	40284	0.702	2.420	3.121	0.923
0.005	149	38284	0.600	2.546	3.147	0.931
0.005	159	36284	0.530	2.687	3.216	0.952
0.005	169	34284	0.463	2.843	3.306	0.978
0.005	179	32284	0.419	3.020	3.438	1.017
0.005	189	30284	0.373	3.219	3.592	1.063
0.005	199	28284	0.336	3.447	3.782	1.119
0.005	204	27284	0.319	3.573	3.892	1.151

Table 8: Cost ratios vs. variance

C_R	n_A	n_E	V_B	V_W	V_T	R_T
0.005	139	40284	0.702	2.420	3.121	0.923
0.01	154	37284	0.563	2.615	3.177	0.940
0.02	189	35534	0.373	2.743	3.117	0.922
0.05	204	36284	0.319	2.687	3.006	0.889
0.1	204	36784	0.319	2.650	2.969	0.878
0.2	204	37034	0.319	2.632	2.951	0.873
0.5	204	37184	0.319	2.622	2.941	0.870

4.6 Maximization of overlap of non-certainty metropolitan areas.

To evaluate the potential gains in expected overlap from using an overlap maximization procedure for selecting new sample PSUs from the group of non-certainty metropolitan areas, we first had to create a new stratification of these areas. We assumed that the number of such strata would be the same as in the current design, 45. The new stratification was created using the procedure of Appendix A, with employment and wage data obtained from the most recent available frame. The old stratification is simply the stratification used to select the current sample of PSUs. Both the old and the new designs are one PSU per stratum, and consequently the selection probability for each PSU is its frame employment divided by the frame employment of its stratum.

There are many procedures for overlap maximization in the literature, most of which are described in Ernst (1999). They are all based on the same general principle, namely, to preserve the unconditional selection probability of a PSU in the new design, but condition its probability of selection on the set of old sample PSUs in such a manner that the conditional probability of selection of a PSU in the new sample is in general greater than its unconditional selection probability when the PSU was in the old sample and less otherwise.

We selected two overlap procedures, those of Perkins (1970) and Ohlsson (1996), to use in evaluating the gains in expected overlap from using an overlap procedure. Both are relatively simple procedures to implement and were developed for one PSU per stratum designs with different stratifications in the old and new designs, exactly the situation we have for the NCS redesign. In addition, Ohlsson's procedure has the desirable property, which most other overlap procedures lack, of preserving the independence of PSU selection from stratum to stratum.

There are other overlap procedures, described in Ernst (1999), using linear programming, which generally yield a higher overlap. These are more complex to program and we did not believe that it was worthwhile to write a program for these procedures for this evaluation. They could be considered for the actual selection of sample PSUs in production.

For both Perkins' and Ohlsson's procedures, it is possible to calculate the expected overlap, that is the expected number of PSUs in common to the old and new samples, directly from the old and new stratifications and probabilities of selection. Ohlsson (1996 Section 3.2) presents an algorithm for computing the expected overlap. Perkins does not, but we provide one in Appendix D.

Over all 45 strata, the expected number of PSUs overlapped using Perkins' method was 26.4, which is an expected proportion of .59. The corresponding numbers for Ohlsson's method were 29.0 and .64. For the selection of the new PSUs without use of an overlap procedure, that is with the new samples selected independently of the old sample PSUs, the corresponding numbers were 15.1 and .34. (The expected overlap in this case was calculated by multiplying the old and new selection probabilities for each

PSU and summing the result over all PSUs.) Thus, the gain in expected overlap from using Ohlsson's procedure over no overlap procedure was 13.9 PSUs

Finally, since neither Perkins' nor Ohlsson's procedure yields an optimal overlap, we calculated an upper bound on the optimal overlap. This was done by taking the minimum of the old and new selection probabilities for each PSU and summing the results over all PSUs. The upper bound was 38.9 or an expected proportion of .87. Thus, the difference between the upper bound and the expected overlap using Ohlsson's procedure was 9.9, indicating that it may be worth exploring the use of other overlap procedures, although we do not know how close we can get to this upper bound with an optimal procedure.

Based on these results a decision has been made to use some overlap procedure to select the new sample non-certainty metropolitan areas.

5. Transition and Future Research

As was mentioned earlier, NCS uses a 5-year rotating panel design with 1/5 of the non-certainty sample units rotating in and out of the sample each year. Therefore, there is going to be a transitional period in which NCS contains sample units from both the old and new area designs. This raises a number of issues about producing locality estimates for areas that were not part of the old design, for old areas that were not selected as part of the new design, and for areas that are in both the new and old design but incurred a definitional change, such as the addition or removal of counties.

There are several transition plans being discussed, all of them with the constraint that collection resources are fixed. Most of the plans call for some sort of rapid implementation of new certainty areas that are of interest to the President's Pay Agent and, consequently, a slow phase in of the new design in other areas. This introduces sampling, weighting, and operational complications. However, no matter which transition plan is used, there is going to be a period of time where some locality and all national-scope estimates will be produced with a mixture of the two area designs.

During the transition period, we will also have the issue of how to estimate variances with the two area designs. NCS uses Fay's variation of balanced repeated replication (BRR) to estimate variances. Any BRR estimate depends on a collection of "replicate" estimates with each replicate estimate being formed by overweighting half of the sample and underweighting the other half. Construction of the set of half-samples first requires the partition of the sample into variance strata. The variance strata are obtained by collapsing together two or more sampling strata, which vary with the design. Therefore, when we have sampled units from the two different area designs, we intend to preserve the variance strata from the old design alongside the variance strata from the new design. This implies the number of replicate estimates will have to be increased to be greater than the total number of variance strata for the two designs.

The area redesign of NCS also gives us the opportunity to revisit other sampling decisions. For example, the old area design involved choosing a sample for each area ignoring census division. Census division estimates were produced by placing all units in an area into only one census division, based on the census division containing the majority of the area's employment. For example, sample establishments in the Connecticut portion of the New York CMSA were included in the Middle Atlantic division for estimation purpose, even though all of Connecticut is actually part of the New England division. In the new design all sample establishments will be included in the correct census division. To facilitate this, we plan to sort on census division within each industry sampling cell for an area, to obtain an implied stratification for census division without adding another formal layer of stratification and further increasing the complexity of the sample design. This should be an improvement on census division estimates of the past.

Other sampling decisions we want to revisit given the opportunity include the following: sample size allocation among the individual certainty areas of interest to the Pay Agent and also the remaining sample areas in aggregate; sample size allocation for industries within an area; and, the number of occupational selections within an establishment. The sample size allocation among areas currently used was actually based on a predecessor of NCS. We would like to see the allocation done based on current NCS data. The sample size allocation among industries within an area for the NCS Wage sample is proportional to aggregate size; we would like to determine if we could lower the average variance for a locality by over-sampling certain industries as we currently do for ECI and EBS. The number of occupational selections within an establishment is either 4, 6, or 8, depending on the size of the establishment. We would like to determine if this is ideal, or would some other occupational selection scheme, especially for larger establishments, be preferable.

Appendix A. Construction of Sampling Strata

In order to calculate between PSU variances, a set of sampling strata for the PSUs was created for each different situation that was examined. These stratifications were formed by following the same general approach used to create the strata for the current area design.

To create a stratification, an overall number of sampling strata was specified for each group of PSUs, such as non-certainty metropolitan PSUs or non-metropolitan PSUs. The stratification then proceeded separately in each group and it is understood that throughout the remainder of this appendix our universe of PSUs is restricted to a single group. Next, the total number of strata within a group was allocated among the nine census divisions proportional to employment. The exact noninteger allocations were then rounded up for the divisions for which the exact allocations had the largest remainder and rounded down for the remaining divisions, with exactly enough allocations rounded up so that the sum of the allocations to the nine census divisions equaled the total allocation. The PSUs in

each census division were then distributed among the calculated number of sampling strata for that division. PSUs were assigned to sampling strata according to the following process:

1. The PSUs in each census division were sorted by a stratification variable (mean wage except in 4.3).
2. The target employment amount for each stratum in a census division was calculated by dividing total employment in the census division by the number of strata in that census division.
3. The PSUs were assigned (following the sorted list) to strata, with the objective of creating strata with employment as close to the target employment as possible. After each PSU was added to a stratum, the employment in that stratum was calculated. When the stratum employment was found to exceed the target employment, then the most recently included PSU was examined. If the stratum's employment with this PSU was closer to the target employment than without it, then that PSU became the last PSU in that stratum. Otherwise, it became the first PSU in the next stratum. Although we did not do this, a stratification with more even employment could have been obtained by recalculating the target employment after each stratum.

Appendix B. Calculation of Between PSU Variance

Between PSU variance was first calculated by selecting 100 independent samples from the frame, and variance was estimated over these samples. However, this method led to variance estimates that were unstable. We tried to estimate the variance with larger numbers of simulated samples, but this did not stabilize the estimates completely, so in order to produce stable estimates, we calculated variance using Taylor series approximations. The Taylor Series formula for the All Workers estimate for mean wages reduces to a calculation of exact variance, since employment, the denominator of the estimate, is a constant for this domain.

The Taylor Series approximation of the variance for mean wages for a domain was calculated using the following formula:

$$\frac{1}{X^2}\sum_{i=strata}\sum_{j=PSU} p_{ij}\left[\frac{(Y_{ij}-X_{ij}R)}{p_{ij}}-(Y_i-X_iR)\right]^2$$

- where X = employment in the domain, Y = wages in the domain, and $R = Y/X$
- Y_i = total wages in the domain in stratum i
- X_i = total employment in the domain in stratum i
- Y_{ij} = total wages for PSU j in stratum i
- X_{ij} = employment for PSU j in stratum i
- p_{ij} = sampling probability of PSU j in stratum i (calculated as the employment in PSU j divided by the employment in stratum i)

Appendix C. Creation of Homogeneous and Heterogeneous Multi-county Outside CBSAs PSUs

The assignment of counties to homogeneous multi-county PSUs was carried out in a manner similar to the process that assigned PSUs to sampling strata described in Appendix A.

The assignment of counties to heterogeneous multi-county PSUs was a bit more complicated. The number, n, of heterogeneous PSUs in each census division was specified as equal to the number of homogeneous PSUs for that census division. The counties were then sorted by wage in descending order. For n heterogeneous multi-county PSUs in a census division, the first n counties in the wage-sorted list were assigned to PSUs 1 through n in order. The PSUs were then sorted in descending order of average wage, and the next n counties in the wage-sorted list were assigned to the newly ordered PSUs in reverse order. After each additional assignment of n counties, the process of sorting the PSUs in descending order of mean wages and assigning the next n counties in reverse order was repeated. When a PSU reached employment equal to or greater than the target employment for its census division, no more counties were assigned to that PSU. This process continued until every county was assigned to a PSU.

Appendix D. Expected Overlap for Perkins' Method

Let S be a stratum in the new design. Let r be the number of strata in the initial design that have a nonempty intersection with S. Let $n_i,\ i=1,...,r$, denote the number of PSUs in the i-th such initial stratum that are in S. Let $p_{ij}, \pi_{ij},\ i=1,...,r,\ \ j=1,...,n_i$, denote the initial and new selection probabilities, respectively, for the j-th PSU in S from initial stratum i. Let

$$y_i = \sum_{j=1}^{n_i}\pi_{ij}$$

Then the probability that PSU ij was in sample for both designs is:

$$\min\left\{y_i\, p_{ij},\ \pi_{ij}\right\}$$

The expected number of overlapping PSUs is obtained by summing this probability over all ij.

6. 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. (1999). The Maximization and Minimization of Sample Overlap Problems: A Half Century of Results. International Statistical Institute, Proceedings, Invited Papers, IASS Topics, 168-182.

Ohlsson, E. (1996). Methods for PPS Size One Sample Coordination. Institute of Actuarial Mathematics and Mathematical Statistics, Stockholm University, No. 194.

Perkins, W. (1970). 1970 CPS Redesign: Proposed Method for Deriving Sample PSU Selection Probabilities Within 1970 NSR Strata. Memo to Joseph Waksberg, U.S. Bureau of the Census.

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