Simulation Study to Validate Sample Allocation for the National Compensation Survey

Hyunshik James Lee[1], Tiandong Li[1], Klaus Teuter[1],
Chester H. Ponikowski[2], and Gwyn R. Ferguson[2]
[1]Westat, 1600 Research Blvd, Rockville, MD 20850
[2]Bureau of Labor Statistics, 2 Massachusetts Ave, NE Washington, DC 20212

Abstract

The National Compensation Survey (NCS) conducted by the Bureau of Labor Statistics (BLS) is an establishment survey which produces as one of its outputs an estimate of the employer costs for employee compensation (ECEC) for establishments in the United States. The survey has been redesigned by reducing sampling stages from three to two, and a new set of sample allocation goals have been derived. This current study is to validate the new allocation goals in comparison with the current sample allocation goals via simulation. The study uses a simulated population dataset, and calculates the population level relative standard errors of the ECEC estimates under the two sets of sample allocation goals using the variance formula for the two-stage PPS design, where the PPS method is used to select establishments at the first stage and to select occupations within the selected establishments at the second stage. This paper discusses these steps and the comparative merits of the two sets of sample allocation goals, demonstrating that the current allocation needs to be revised.

Key Words: Replicated population, random rounding of the weighting factor, Yates-Grundy-Sen variance formula, approximate variance formula for PPS designs.

1. Introduction

The National Compensation Survey (NCS) is an establishment survey conducted by the Bureau of Labor Statistics (BLS) which provides comprehensive measures of employer costs for employee compensation (ECEC), compensation trends, and the incidence and provisions of employer-provided benefits. The survey covers all workers in private industry establishments and in State and local government, in the 50 States and the District of Columbia. Establishments with one or more workers are included in the survey. Excluded from the survey are workers in the Federal Government and quasi-Federal agencies, military personnel, agricultural industry, workers in private households, the self-employed, volunteers, unpaid workers, individuals receiving long-term disability compensation, individuals working overseas, individuals who set their own pay (for example, proprietors, owners, major stockholders, and partners in unincorporated firms), and those paid token wages.

The BLS Quarterly Census of Employment and Wages (QCEW) serves as the sampling frame for the NCS survey. The QCEW is created from State Unemployment Insurance (UI) files of establishments, which are obtained through the cooperation of the individual state agencies (BLS Handbook of Methods, Chapter 5).

Recently the NCS has undergone a sample redesign. The redesigned NCS sample consists of three rotating replacement sample panels for private industry establishments, an additional sample panel for State and local government entities, and an additional panel for private industry firms in the aircraft manufacturing industry. Each of the sample panels is in the sample for at least three years before it is replaced by a new sample panel selected annually from the most current frame. Establishments in each sample panel are initiated over a one-year time period. After initiation, data are updated quarterly for each selected establishment and occupation until the panel in which the establishment was selected is replaced.

The transition to the redesigned sample started in the spring of 2012 with the fielding of the first private industry sample and will continue until late 2016 when the State and local government sample enters. As a part of the continuous process of survey improvement, establishment sample allocation is being studied to determine if adjustments to the current sample allocation could result in even more precise survey estimates.

In the previous research of sample allocation for the redesigned NCS, new sample allocation was obtained for the private industry strata in the new design using variance components, design effect, and response rates estimated from the three-stage NCS sample prior to redesign (see Lee et al., 2012 for detail). A Monte Carlo simulation study was conducted to compare the current (old) allocation with the new allocation in terms of the relative standard error of the NCS major estimate, the employer costs for employee compensation (ECEC).

This paper is organized as follows: Section 2 provides some literature review and describes the data source and the method of generating the simulation population. Section 3 explains the NCS sample design, sample allocation to the sampling strata, identification of certainty units, and selection of 1,000 simulated samples. Section 4 discusses the variance components of the ECEC estimate due to two-stage sampling, methods of calculating the components from simulated samples and by theoretical formula, and computation of simulated and theoretical RSEs. Section 5 provides comparison of the simulation results for two allocation goals and comparison of the simulation results. It also discusses the question of simulation errors and comparison of the simulation results with theoretical results obtained from population level formulae. Section 7 presents some concluding remarks, some caveats, and recommendations.

2. Creation of the Population Data for Simulation

For the study, we needed a population frame of all private establishments with data for compensation and employment at the occupation level under the NCS coverage. Since there is no such data available, we generated one from the Occupational Employment Statistics (OES) sample data, which uses the same sampling frame as NCS but has a much larger sample size (1.2 million) than NCS (about 10 thousands). The large sample size provides a huge advantage for generation of population data.

The generation approach is based on the bootstrap idea of Efron (1981), which is widely used for various purposes. Efron's original idea was to obtain the sampling distribution of a statistic of an IID (independently identically distributed) sample. Gross (1980) first adapted the bootstrap idea in survey sampling for a survey design that is the stratified simple random sampling (SRS) of clusters. The idea was to create a pseudo or synthetic

population by replicating each sample unit (cluster) by its weighting factor and then to select many bootstrap samples using the original sample design (SRS) to estimate the variance of the sample median. Our approach resembles this idea, where a synthetic population dataset is generated by replicating OES sample units by the integerized weighting factors with random rounding. However, we use the NCS design to select simulated samples instead of the OES design, which would be used by Gross.

Generation of synthetic populations from a sample data was also proposed by Raghunathan, Reiter, and Rubin (2003) in a different context. Their main purpose was to protect the respondent data from disclosure by generating multiple sets of the population data and selecting multiple sets of analysis data from the synthetic population datasets using SRS, and these samples are used to estimate parameters of interests and their variances. Their approach to generate the synthetic population data is based on the multiple imputation method to impute unobserved units in the population. The usual multiple imputation method uses extensive modeling to obtain a regression model under the posterior distribution and then imputes unobserved values using the model, which we want to avoid. However, there is another way of generating a synthetic population, that is, the Bayesian Bootstrap method proposed by Schenker and Rubin (1986), which was also used by Raghunathan et al. (2003). This involves random sampling of observed units instead of model development and prediction.

After considering various options based on the basic methodologies discussed above, we adopted the idea of Gross (1980) to generate the population dataset for simulation because it would serve our purpose well with a simpler method and is based on the preferred design-based framework.

2.1 Brief Description of the OES Sample Data

The full OES sample data consisting of 6 panels represent the full universe with about 9,526,500 occupation level records (about 8 occupational records per establishment). The OES uses the stratified design and selects about 200,000 establishments for each panel by probability proportional to the estimated size (PPeS) sampling method from each non-certainty stratum cell after identifying certainty units with a large size (usually 250+ employees). The estimated size for the noncertainty units is calculated as the average number of establishments' total employees for each of the 6 noncertainty size classes (1-4, 5-9, 10-19, 20-49, 50-99, and 100-249) within the state. There is no subsampling of occupations, so the establishment weights are the same as the occupation weights. Due to the sampling method that favors large units, the OES sample accounts for about 66 percent of the total employment, while the sample size (of 1.2 million) is roughly one fifth of the universe.

The response rate of the OES survey in the 6-panel combined sample in May 2011 was about 77 percent, and OES uses imputation for nonrespondents. The OES data we used contains about 1,120,300 viable establishments (that are not out of scope or out of business). For a sampled establishment, wages data including salaries are collected for all occupations (i.e., there is no subsampling of occupations as done for NCS) but the wage data are collected in terms of employment distribution of the 12 preset wage intervals. OES does not use the wage intervals directly to estimate the total wages but uses the NCS data to calculate the mean hourly wage rate for each interval by ECI (Employment Cost Index) occupation group (A-H, J, K, X). Further, wages data for each panel except the current are adjusted by the aging factor that varies by panel year and ECI occupation group. The total wages are then calculated for each occupation in an establishment as the

sum of 12 terms of the mean hourly wage rate multiplied by the number of employees for the interval. Therefore, the OES data is at the occupation level, not at the employee level. This has an important implication, which will be discussed further later.

2.2 Method to Generate the Population Data for Simulation

The population data for simulation was generated by replicating the OES sample establishments by their sampling weights. A non-integer weight was randomly rounded up to the smallest integer larger than the weight with a probability of a and down to the largest integer smaller than the weight with a probability of $(1-a)$, where a is the fractional part of the weight. The generated population size (denoted by N) is random but equal to the OES population size in expectation.

We examined whether multiple synthetic population datasets would be needed but it was found that the ECEC NCS produces as the main statistic varies only slightly among synthetic populations that would be generated through replication by randomly rounded sampling weights. Therefore, only one synthetic population was generated for simulation. There were about 5,897,000 establishments in the generated population, which is slightly lower than the weighted count (about 6,027,700) by OES sampling weights.

2.3 Statistics for the Generated Population

Table 2.1 shows the population size of establishments and ECEC by industry (detailed and aggregated) for the generated population, where five aggregated industries are defined by grouping detailed industries as shown by the first digit of the detailed industry code (see Table 2.1). The population sizes are rounded to 100 for confidentiality reason. The ECEC values by detailed industry range from 10.94 for "Accommodation and Food Services" to 34.18 for "Professional, Scientific, and Tech Services". These ECECs are of the generated population, and generally different from those published by BLS based on samples.

Table 2.1: Population size and ECEC by detailed and aggregated industries

Detailed Industry		*Pop Size*[1]	*ECEC*[2]
1 - 21A	Mining	24,000	28.11
1 - 23A	Construction	540,500	23.22
1 - 31A	Manufacturing	293,400	22.45
2 - 52A	Finance (excluding Insurance)	220,000	29.29
2 - 52B	Insurance Carriers and Related Activities	146,800	28.57
2 - 53A	Real Estate and Rental and Leasing	268,800	19.81
3 - 61A	Education	49,700	22.38
3 - 61B	Elementary & Secondary Education	72,300	22.05
3 - 61C	Colleges & Universities	8,300	28.36
4 - 62A	Health and Social Assistance	603,300	23.23
4 - 62B	Hospitals	9,500	26.42
4 - 62C	Nursing Homes	62,600	15.25
5 - 22A	Utilities	17,800	31.4
5 - 42A	Wholesale Trade	393,000	24.54
5 - 44A	Retail Trade	865,900	14.25
5 - 48A	Transportation and Warehousing	166,600	21.37
5 - 51A	Information	106,700	29.96
5 - 54A	Professional, Scientific, and Tech Services	666,500	34.18
5 - 55A	Management of Companies and Enterprises	36,100	33.32

Table 2.1: Population size and ECEC by detailed and aggregated industries (Continued)

Detailed Industry		*Pop Size*[1]	*ECEC*[2]
5 - 56A	Admin and Support, Waste Management	315,600	16.54
5 - 71A	Arts, Entertainment, and Recreation	88,700	15.68
5 - 72A	Accommodation and Food Services	485,500	10.94
5 - 81A	Other services except public administration	455,400	17.39
Aggregated		*Pop Size*[1]	*ECEC*[2]
1	Good Producing	857,900	22.9
2	Finance, Insurance, and Real Estate	635,700	26.75
3	Education	130,200	23.87
4	Health Care, including Hospitals and Nursing Care	675,400	22.88
5	Service Providing	3,597,700	19.46
National		5,897,000	21.46

[1] The population sizes are rounded to 100, and the rounded aggregated numbers may not be the same as the sum of rounded numbers.

[2] The ECECs in the table were obtained from the generated population data, and therefore, are different from those in the BLS publications.

3. Selection of Simulation Samples

To select simulation samples, the NCS sample design was imposed on the generated population explained in the previous section. Then two sample allocation goals were applied to determine the sample size at the design strata. The sample allocation goals are given at the detailed industry level but the detailed industry is not a stratification variable. Detailed industry level sample allocation was achieved through modification of establishment's employment to be used as the measure of size (MOS) for PPS sampling. Then 1,000 simulated samples were selected for each allocation goal. This process is explained in this section.

3.1 NCS Sample Design

The redesigned NCS sample is selected using a two stage stratified design, where strata are defined by 24 geographic areas crossed by 5 aggregated industries (120 strata altogether). The 24 areas consist of the 15 largest metropolitan areas by employment and the rest of each of the nine Census Divisions, excluding the 15 largest metropolitan areas.

From each of the 120 sampling strata, a systematic PPS (probability proportionate to employment size) sample of establishments using the modified employment size as the measure of size (MOS) is selected with the sample size determined as described below. The sample size at the sampling stratum level is determined by allocating the national sample sizes for 23 detailed industries as shown in Table 3.1 to the sampling strata. Although the national sample sizes are allocated at the detailed industry level, industry stratification is not at the detailed level but at the aggregated level. Therefore, these detailed industries are substrata without hard stratum boundaries within sampling strata. In addition, when the national aggregated industry sample size is allocated to sampling areas within the aggregated industry, the allocation is proportional to the area total employment of establishments belonging to the aggregated industry. However, simple proportional allocation would not ensure the allocated sample size for the detailed industry as the employment distribution across the detailed industries within each sample stratum differs from the detailed industry target sampling percentages (see Table 3.1). To address this issue, the original MOS based on the employment size is modified by

multiplying the employment size by an MOS adjustment factor, which is defined as the ratio of the target percentage to the percent employment distribution for each detailed industry substratum. Using this modified employment size as the MOS, allocation to the aggregated industry level is performed. During this process, certainties are identified (see Table 3.1). This is a rather complex iterative procedure (for details, refer to Lee and Li, 2013).

The second stage is a probability selection of occupations within the establishments. A more detailed description of the new NCS sample design is given in Ferguson, et al. (2010) while a description of the estimates produced and the estimation methodology is given in Chapter 8 of BLS Handbook of Methods.

Table 3.1: Total sample size and the number of certainties at the detailed industry level for current and proposed allocations for private industry

Industry	*Sample Size*		*Percent Distribution*[1]		*Certainty Selection*	
Detailed	*Curr*	*Prop*	*Curr*	*Prop*	*Curr*	*Prop*
1 - 21A	90	556	4.5	30.2	0	55
1 - 23A	912	852	45.1	46.3	0	0
1 - 31A	1,020	431	50.4	23.4	6	0
2 - 52A	954	1,000	54.7	64.2	15	16
2 - 52B	581	168	33.3	10.8	9	0
2 - 53A	208	390	11.9	25	0	0
3 - 61A	58	289	13.8	41.8	0	3
3 - 61B	88	235	21	34	1	6
3 - 61C	274	167	65.2	24.1	26	9
4 - 62A	186	533	22.5	70.7	0	1
4 - 62B	254	115	30.7	15.3	0	0
4 - 62C	387	106	46.8	14	0	0
5 - 22A	120	102	2.5	2.1	2	2
5 - 42A	702	539	14.8	11	0	0
5 - 44A	1,448	154	30.5	3.1	0	0
5 - 48A	315	1,000	6.6	20.4	3	60
5 - 51A	370	201	7.8	4.1	1	1
5 - 54A	408	604	8.6	12.3	0	0
5 - 55A	70	322	1.5	6.6	0	3
5 - 56A	458	616	9.7	12.5	2	2
5 - 71A	106	190	2.2	3.9	1	2
5 - 72A	391	185	8.2	3.8	0	0
5 - 81A	354	1,000	7.5	20.4	0	0
Aggregated						
1	2,022	1,838			6	55
2	1,743	1,558			24	16
3	420	691			27	18
4	827	753			0	1
5	4,742	4,914			9	70
National	9,754	9,754			66	160

[1] Within aggregated industry

As mentioned in Section 1, this study compares the current (old) allocation with the proposed (new) allocation for private industry, excluding aircraft manufacturing. The

sample sizes in Table 3.1 show these sample allocation goals both of which have a total sample size of 9,754.

The current allocation is the sample allocation that is currently being used by BLS, so it is used as the baseline allocation to compare with the proposed allocation. The proposed allocation is based on the previous sample allocation research (Lee et al., 2012), but adjusted by imposing a cap of 1,000 to each detailed industry and redistribute the excess sample size proportionally to other industries, keeping the overall sample size at 9,754.

3.2 Sample Selection

For each sample allocation, identified certainties are first removed from the sampling frame, and then using the SAS SURVEYSELECT procedure with the modified sample size and MOS, a non-certainty sample is selected from each sampling cell. The sampling method is systematic PPS sampling with a sorted sampling frame by the two sort variables (detailed industry and MOS) within each sampling stratum. Repeating this process, we selected 1,000 independent samples of non-certainties for each allocation using a different random seed each time. The target sample sizes as shown in Table 3.1 have been achieved for both allocation goals.

As explained before, we did not have employee level records in the generated population data, and thus, there was no simulation for subsampling of occupations. Therefore, within-establishment variance was calculated theoretically as explained later.

4. Variance Components and RSE of the ECEC Estimate

We compared two allocations in terms of the relative standard error (RSE) of the employer cost for employee compensation (ECEC), where RSE is defined as the ratio of the (sampling) standard error of the ECEC estimate to the estimate. The variance of the ECEC estimate, which is needed to calculate RSE, has two components: (1) the first-stage component (between establishments); and (2) the second-stage component (within establishments). These variance components themselves were of interest, but the main statistic to compare was RSE. Variance components and RSE were obtained using 1,000 simulation samples.

It should be noted that the ECEC estimate used in this study is different from what is actually used by NCS because the generated population data we used for simulation do not have total compensation including benefits but just wages and salaries. The simulation population has been developed from the OES sample data, whose occupation-level data structure within the establishment is different from the real world structure from which the actual NCS sample is selected. The field worker who selects the occupation sample uses the list of employees provided by the establishment without modification. It can be ordered in many different ways we do not know. For this reason, mimicking the NCS sampling procedure as done in the field is extremely difficult or almost impossible in simulation. Furthermore, the generated population data lack employee level data. Therefore, within-establishment sampling variance was calculated using a formula as described in Section 4.3 assuming the list is randomly ordered and using the network sampling approach.

4.1 Variance Components for the ECEC Estimate

The ECEC estimate is a ratio of the per hour total wages and salaries to the total number of employees. Denoting the population ECEC by R, it is defined by:

$$R = Z/Y = \sum_{i=1}^{N}\sum_{j=1}^{M_i} z_{ij} / \sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij} \tag{4.1}$$

where z_{ij} is occupational wages and salaries and y_{ij} is occupational employment for occupation j in establishment i, N is the total number of establishments in the population frame, M_i is the number of occupations in establishment i, $Z = \sum_{i=1}^{N}\sum_{j=1}^{M_i} z_{ij}$, and $Y = \sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}$. From a sample of establishments and subsamples of occupations from the sampled establishments, the population ECEC is estimated by a ratio estimate:

$$\hat{R} = \hat{Z}/\hat{Y} \tag{4.2}$$

where $\hat{Z} = \sum_{i=1}^{n} \pi_i^{-1} \sum_{j=1}^{m_i} \pi_{j|i}^{-1} z_{ij}$, $\hat{Y} = \sum_{i=1}^{n} \pi_i^{-1} \sum_{j=1}^{m_i} \pi_{j|i}^{-1} y_{ij}$, π_i is the selection probability for establishment i, $\pi_{j|i}$ is the probability of selecting occupation j given establishment i selected, n is the establishment sample size, m_i is the occupation sample size for establishment i.

From the theory of ratio estimation, we can obtain an approximate variance of $\hat{R}$ using the linearized value. Let $t_{ij} = (z_{ij} - R y_{ij})/Y$, $T_i = \sum_{j=1}^{M_i} t_{ij}$, and $\hat{T}_i = \sum_{j=1}^{m_i} w_{j|i} t_{ij}$ with $w_{j|i} = \pi_{j|i}^{-1}$. An approximate variance of $\hat{R}$ is given by:

$$V(\hat{R}) \cong V(\hat{T}) \tag{4.3}$$

where $\hat{T} = \sum_{i=1}^{n} \hat{T}_i/\pi_i$ and $\hat{T}_i = \sum_{j=1}^{m_i} \pi_{j|i}^{-1} t_{ij}$. Note that we use the population R, not an estimate, which is available from the population data, in the definition of the linearized value. Using the true R, the approximation in (4.3) becomes very good, and the bias is negligible.

Then by the usual variance component formula, we can write (ignoring the stratification for now for ease of notation and using the usual notation):

$$\begin{aligned} V(\hat{T}) &= V_1 E_2\left(\sum_{i=1}^{n} \hat{T}_i/\pi_i\right) + E_1 V_2\left(\sum_{i=1}^{n} \hat{T}_i/\pi_i\right) \\ &= V_1\left(\sum_{i=1}^{n} T_i/\pi_i\right) + E_1\left(\sum_{i=1}^{n} V_2(\hat{T}_i)/\pi_i^2\right) \\ &= V_1\left(\sum_{i=1}^{n} T_i/\pi_i\right) + \sum_{i=1}^{N} V_2(\hat{T}_i)/\pi_i \end{aligned} \tag{4.4}$$

where subscripts 1 and 2 indicate the first- and second-stage variance and expectation, respectively, under the new NCS design. The first and second terms in the above expression are, respectively, the first and second variance components of the total variance, which are also referred to as the between-PSU (between establishments) variance and the within-PSU variance. Note that breaking up the total variance into the two variance components as shown in (4.4) was possible because of the linearization.

Using the Yates-Grundy-Sen (YGS) formula (Yates and Grundy, 1953; Sen, 1953), the first component is given as:

$$C_1 = V_1(\sum_{i=1}^{n} T_i/\pi_i) = \sum_{h=1}^{H} V_{YGSh1}\left(\sum_{i=1}^{n_h} T_{hi}/\pi_{hi}\right) \\ = \sum_{h=1}^{H}\left\{\frac{1}{2}\sum_{i=1}^{N_h}\sum_{k=1,k\neq i}^{N_h}(\pi_{hi}\pi_{hk} - \pi_{hik})(T_{hi}/\pi_{hi} - T_{hk}/\pi_{hk})^2\right\} \quad (4.5)$$

where π_{hik} is the joint probability that establishments i and k appear in the sample jointly, H is the total number of sampling strata (i.e., $H = 120$), and $n = \sum_{h=1}^{H} n_h$. In (4.6) all notation has the same meaning as before except they are defined within stratum h.

For the second component, we do not need to calculate it stratum by stratum as it is given by assuming that the sampling method is PPS:

$$C_2 = \sum_{i=1}^{N} V_2(\hat{T}_i)/\pi_i = \sum_{i=1}^{N} V_{YGS2}(\hat{T}_i)/\pi_i \\ = \frac{1}{2}\sum_{i=1}^{N}\pi_i^{-1}\sum_{j=1}^{M_i}\sum_{j'=1,j'\neq j}^{M_i}(\pi_{j|i}\pi_{j'|i} - \pi_{jj'|i})\left(t_{ij}/\pi_{j|i} - t_{ij'}/\pi_{j'|i}\right)^2. \quad (4.6)$$

Note that certainty strata do not contribute any variance in (4.5). We cannot implement (4.6) because of the particular sampling procedure NCS uses to select occupation, which will be explained later.

4.2 Estimation of the First Variance Component

For each simulated sample g, $\sum_{i=1}^{n} T_i/\pi_i$ is computed, which is denoted by $\hat{T}(g)$ for $g = 1,2,3,\ldots,1000$. Then the Monte Carlo (simulated) variance of $\hat{T}(g)$'s is the first-stage variance component by the simulation method. Namely,

$$\hat{C}_{M1} = \hat{V}_{M1}(\hat{T}) = \sum_{g=1}^{1000}\left(\hat{T}(g) - \sum_{h=1}^{1000}\hat{T}(h)/1000\right)^2/999. \quad (4.7)$$

Without simulation, we can calculate the true first variance component using the YGS formula in (4.6) but we need the joint probabilities π_{hik}, which are not easy to compute for the systematic PPS sampling. There is an approximate formula proposed by Berger (2004), which does not require the joint probability for PPS designs that satisfy a certain condition. However, the systematic PPS method NCS uses to select establishments does not satisfy the condition. Pinciaro (1978) provided a computer algorithm that calculates joint probabilities for the systematic PPS design, by which we can compute the needed joint probabilities for (4.6).

4.3 Estimation of the Second Variance Component

The NCS design for selection of occupations from a sample establishment is not an ordinary sampling method as explained below:

1) A systematic sample of employees is selected from the list of all employees given by the establishment;
2) Occupations of selected employees are identified, and wages data for all employees with the identified occupations are obtained and captured;
3) The sample size of the employee sample in 1) is determined according to the occupation selection schedule, and the number of occupations actually selected is less than or equal to this sample size due to possible multiple hits of the same occupation;
4) The employee sampling weight is the total number of employees in the establishment divided by the sample size of the employee sample in 1).

The sampling method described above is one of the network sampling methods (Johnson, 1995). For establishment i, let m_i be the sample size of occupations to be selected according to the occupation selection schedule, L_{ij} be the total employment in occupation j, M_i be the number of occupations, and $L_i = \sum_{j=1}^{M_i} L_{ij}$ be the total number of employees. It should be noted that even though it is intended to select m_i unique occupations, the selection procedure described above does not guarantee this because when m_i employees are selected, more than one employee can be selected from the same occupation.

In any case, the employee sampling weight for sample employee l is given by $\omega_{il} = L_i/m_i$ for $l = 1, 2, \ldots, m_i$, so it is the same for all sample employees. Then the network sampling weight for the j-th selected occupation from establishment i through the sample of employees is given by:

$$w_{ij} = \sum_{l \in j} \omega_{il}/L_{ij} = \sum_{l \in j} L_i/(m_i L_{ij}) = h_{ij}\, L_i/(m_i L_{ij}) \qquad (4.8)$$

where h_{ij} is the number of hits or the number of employees selected from occupation j. The terms in (4.8) are summed over h_{ij} employees belonging to occupation j (see Johnson, 1995). Since the same occupation can be selected more than once, $h_{ij} \geq 1$, and thus, $m_i \geq m_i'$, where m_i' is the number of unique occupations selected. Note that the employee sampling weight is modified by dividing it by the employment size of the occupation to calculate the occupation selection weight. This can be viewed as an adjustment for the fact that the occupation selection probability is proportional to the employment size in "expectation" in repeated sampling. This aspect is an important point that differs from the randomized PPS sampling previously considered. Under the network sampling approach, the certainty issue does not pose any problem in calculation of the weight in (4.8), and this weight provides an unbiased estimate for the total (see Johnson, 1995). For example, $\hat{T}_i = \sum_{j=1}^{m_i'} w_{ij} t_{ij}$ is an unbiased estimator for $T_i = \sum_{j=1}^{M_i} t_{ij}$. We can rewrite this estimator as the weighted sum of the employee sample rather than the occupation sample as follows:

$$\hat{T}_i = \sum_{j=1}^{m_i'} w_{ij} t_{ij} = \sum_{j=1}^{m_i'} \left\{\frac{h_{ij} L_i}{m_i L_{ij}}\right\} t_{ij} = \sum_{l=1}^{m_i} \left(\frac{L_i}{m_i}\right)\left(\frac{t_{il}}{L_{il}}\right) = \sum_{l=1}^{m_i} \omega_{il} U_{il} \qquad (4.9)$$

where $U_{il} = t_{il}/L_{il}$. Note that in (4.9), each term for a multiply hit occupation (i.e., $h_{ij} > 1$) is written out, if it exists, in terms of sample employees. Since $m_i \geq m_i'$, there can be more terms in the summation in the far right expression than that in the far-left expression. We define t-values and L-values at the employee level by setting $t_{il} = t_{ij}$ and $L_{il} = L_{ij}$ for sample employee l in sample occupation j. Note that U_{il} is the average t-values in occupation j for $l \in j$, and the same for all employees in that occupation.

If it is reasonable to assume that the list of employees provided by the sample establishment for selection of occupations is randomly ordered with respect to occupations, it is reasonable to treat the employee sample as an SRS from the employee list. Then $\hat{T}_i$ in (4.9) is an estimator for the total from an SRS of m_i employees, and therefore, the second-stage variance can be obtained from:

$$C_2 = \sum_{i=1}^{N} V_2(\hat{T}_i)/\pi_i = \sum_{i=1}^{N} \pi_i^{-1} \sum_{l=1}^{L_i} \frac{L_i^2(1 - m_i/L_i)}{m_i(L_i - 1)} (U_{il} - \bar{U}_i)^2 = \sum_{i=1}^{N} Q_i \qquad (4.10)$$

where $\overline{U}_i = \sum_{l=1}^{L_i} U_{il}/L_i$ and $Q_i = \pi_i^{-1} \sum_{l=1}^{L_i} \frac{L_i^2(1-m_i/L_i)}{m_i(L_i-1)} (U_{il} - \overline{U}_i)^2$.

We can calculate (4.10) because we have the population data. If the first (variance) component is obtained using the theoretical method given by (4.6), it would be appropriate to use the theoretical (4.10) computed from the population data to obtain the total variance. However, the first component will be computed using simulated samples, so it is better to estimate the population quantity in (4.10) from simulated samples as well. Since the quantity in (4.10) is a population total, we can estimate it by the Horvitz-Thompson estimator from a sample as follows:

$$\hat{C}_2 = \sum_{i=1}^{n} \pi_i^{-1} Q_i = \sum_{i=1}^{n} \pi_i^{-2} \sum_{l=1}^{L_i} \frac{L_i^2(1-m_i/L_i)}{m_i(L_i-1)} (U_{il} - \overline{U}_i)^2 \tag{4.11}$$

Note that the right-hand side of (4.11) is a summation over the (single) sample. Then the Monte Carlo estimate of the second component is given by the average of estimates in (4.11) over 1,000 simulated samples. Namely,

$$\hat{C}_{M2} = \frac{1}{1000} \sum_{g=1}^{1000} \sum_{i \in S_g} \pi_i^{-2} \sum_{l=1}^{L_i} \frac{L_i^2(1-m_i/L_i)}{m_i(L_i-1)} (U_{il} - \overline{U}_i)^2 \tag{4.12}$$

where S_g is the g-th simulated sample.

4.4 Monte Carlo Estimate of the Population RSE of the ECEC Estimate

The population RSE of the ECEC estimate defined as $\text{RSE}(\hat{R}) = V(\hat{R})^{1/2}/R$ is estimated by plugging Monte Carlo estimate for $V(\hat{R})$ given by:

$$\hat{V}_M(\hat{T}) = \hat{C}_{M1} + \hat{C}_{M2} \tag{4.13}$$

and the Monte Carlo simulated average of ECEC estimates for R given by:

$$\hat{R}_M = \frac{1}{1000} \sum_{g=1}^{1000} \tilde{R}_g \tag{4.14}$$

where $\tilde{R}_g = \sum_{i=1}^{n} (Z_i/\pi_i) / \sum_{i=1}^{n} (Y_i/\pi_i)$ computed from simulated sample g. Note that this is different from $\hat{R}$ in (4.2), where occupational sample estimates (i.e., $\hat{Z}_i$ and $\hat{Y}_i$) from sample establishments are involved, whereas establishment level true values (i.e., Z_i and Y_i) are used in $\tilde{R}_g$. Since we do not subsample occupations from each sample establishment to select an occupation sample, we cannot compute $\hat{Z}$ and $\hat{Y}$ using $\hat{Z}_i$ and $\hat{Y}_i$ in (4.2).

Then the Monte Carlo estimate of the population RSE is given by:

$$\widehat{\text{RSE}}_M(\hat{R}) = \sqrt{\hat{V}_M(\hat{R})}/\hat{R}_M = \sqrt{\hat{V}_M(\hat{T})}/\hat{R}_M. \tag{4.15}$$

We could use R, which can be calculated from the population data but it would be better to use the Monte Carlo estimate because the total variance is obtained from the Monte Carlo simulation. For comparison, we also computed true RSE and true variances using

(4.6), (4.10), and the true R. Any deviation of the simulated RSEs from the true RSEs are due to simulation error.

5. Comparison of Simulation and Research Results

As originally planned, RSEs obtained from simulation for two allocations are compared using the percent relative difference between the simulated RSEs for the current and those for the proposed allocation with the current RSE as the base.

The comparison results are presented in Table 5.1. The percent relative difference is defined as $c = 100 \times (a - b)/b$, where b is the base RSE and a is the RSE to be compared. There is no good reason to choose one particular base, and the relative difference can also be computed using the other RSE as the base. So if we change the base for the comparison, the positive relative difference becomes negative, and the absolute magnitude is not the same. If a relative percent difference is c percent, the opposite based percent relative difference (c') is given by $c' = -c/(100 + c)$. Therefore, if $c > 50$, then $c' < -100/3 = -33.\dot{3}$ with the limiting value of -100.

Table 5.1: Comparison of RSEs for the current and proposed allocations

Level	*Current*		*Proposed*		*Rel Diff (%)*
Detailed	*Samp Size*	*RSE (%)*	*Samp Size*	*RSE (%)*	*Prop vs. Curr*
1 - 21A	90	13.09	556	4.05	-69.04
1 - 23A	912	1.39	852	1.62	16.21
1 - 31A	1,020	1.59	431	2.13	34.43
2 - 52A	954	2	1,000	2.11	5.18
2 - 52B	581	3.87	168	3.78	-2.41
2 - 53A	208	4.11	390	4.01	-2.51
3 - 61A	58	16.31	289	6.14	-62.39
3 - 61B	88	2.37	235	1.39	-41.28
3 - 61C	274	1.68	167	2.17	29.09
4 - 62A	186	5.05	533	3.11	-38.49
4 - 62B	254	3.41	115	2.14	-37.2
4 - 62C	387	1.63	106	3.85	135.78
5 - 22A	120	2.29	102	2.44	6.64
5 - 42A	702	2.04	539	2.97	45.5
5 - 44A	1,448	2.78	154	3.71	33.64
5 - 48A	315	4.23	1,000	1.92	-54.61
5 - 51A	370	2.82	201	3.15	11.58
5 - 54A	408	2.64	604	2.28	-13.56
5 - 55A	70	4.83	322	2.73	-43.46
5 - 56A	458	2.8	616	2.99	7.11
5 - 71A	106	6.11	190	4.41	-27.89
5 - 72A	391	1.63	185	2.35	44.62
5 - 81A	354	3.02	1,000	1.8	-40.58
Aggregated					
1	2,022	1.19	1,838	1.47	23.47
2	1,743	1.6	1,558	1.68	5.01
3	420	1.97	691	1.26	-36.01
4	827	2.7	753	1.94	-28.01
5	4,742	0.94	4,914	1.1	16.46
National	9,754	0.72	9,754	0.7	-2.6

Note: RSEs larger than 5 percent are highlighted in yellow, and the percent relative difference larger than 50 percent or less than -33.3 percent are also highlighted in orange.

As a summary of what is presented in Table 5.1, some summary statistics of the detailed industry level RSEs are shown in Table 5.2. We consider the allocation that gives a smaller mean with a smaller variation (standard deviation) better. The average RSE under the current allocation is 37 percent larger than the average RSE under the proposed, and the variability of the RSEs for the current allocation is much larger than those of the proposed. This is expected because the current allocation was not designed to have the same RSE by detail industry. However, the RSE for some industries such as 61A and 21A is clearly out of acceptable limit.

Table 5.2: Summary statistics for detailed industry level percent RSEs

Statistic	*Minimum*	*Maximum*	*Median*	*Mean*	*Stand Dev*
Current	1.39	16.31	2.8	3.99	3.63
Proposed	1.39	6.14	2.44	2.92	1.11

Comparison at the aggregated industry level largely reflects that of the detailed industry level comparison. However, the RSEs at the national level are quite comparable, where the current national RSE is slightly larger than the proposed one. Therefore, different allocation methods at the detailed industry level have considerable impact for the detailed industry RSEs but the differences are virtually wiped out at the national level.

Extreme RSEs (16.31 percent for 61A and 13.09 percent for 21A) under the current allocation are associated with small sample sizes. No outliers were found for the proposed allocation. Such a sharp contrast can be explained by radically different apportionment to those industries by the two allocations; the current allocation assigns a sample sizes 90 and 58 to 21A and 61A, respectively, compared to 556 and 289 under the proposed allocation. Industry 21A has many large establishments since 55 establishments were selected as certainty under the proposed allocation, while no certainty was selected under the current. There are some large units in industry 61A as well but not many. However, whether those few large ones are selected as certainty or not makes a huge difference in the variance because no extreme variance was observed under the proposed allocation as the few large establishments were selected as certainty under the proposed allocation (see Table 3.1 for the number of certainties). Therefore, it is very important to allocate an adequate sample size so that large establishments are selected with certainty to avoid an extreme variance.

The second variance component is a much smaller part in the total variance and was not a cause to the extreme variance.

The theoretical variance component formulae given in (4.6) and (4.10) and RSEs can be calculated using the population data. We tried to use them with the Pinciaro algorithm (1978) to calculate joint probabilities needed in (4.6). The algorithm is very computer-intensive, especially when the stratum size is large (e.g., one of the sampling strata has nearly a half million establishments), but provides accurate joint probabilities. Comparison of the simulation and theoretical results indicates that the simulation size of 1,000 was not nearly enough.

6. Conclusions and Recommendations

This study was conducted to validate the proposed allocation that resulted from the previous sample allocation research using simulation. The main thrust was to compare the performance of the proposed allocation with the current allocation.

The population data for simulation were generated from the OES sample data. The redesigned NCS design was implemented on the sampling frame of the generated population data. Simulation was conducted by selecting 1,000 independent samples of establishments with sample sizes determined for the sampling strata for each of the two sample allocation goals (the current and proposed). The second-stage sampling (i.e., subsampling of occupations) was not carried out due to unavailability of employee level data in the generated population data and for the difficulty to mimic the field procedure to select the occupation sample. From these simulated samples, the Monte Carlo (simulated) RSEs were obtained at the detailed and some aggregated industry levels. The simulation results were compared for the two allocations.

Even with the simulation size of 1,000 samples, the simulation errors can be substantial, especially for some detailed industries with large variance. Nevertheless, the comparison of the current and proposed allocations is fair, and it reveals that the proposed allocation performs considerably better than the current allocation in terms of the average of detailed industry level RSEs and their standard deviation. As with any study such as this, there are some caveats to be noted:

1. The ECEC definition used in the study is different from that used by NCS as the total compensation used in the NCS definition includes wages, salaries, and benefits, whereas benefits are not included in the study definition.
2. The simulation results are subject to two sources of errors, simulation error and error involved in the generation of the population data, which were generated from the OES sample data. It was shown that the simulation error was substantial.
3. The second stage variance component could not be fully simulated because employee level occupational data were not available.

It seems obvious that the current allocation grossly under-allocates industry 21A and 61A since the RSEs for the two industries under the current allocation are unusually high. Therefore, it is recommended to substantially increase the sample size for these industries. Furthermore, despite the caveats mentioned above, we believe that the current or proposed allocation can be improved by using the results presented in this study.

References

Berger, Y.G. (2004). A simple variance estimator for unequal probability sampling without replacement. *Applied Statistics*, 31, 305-315.

Efron, B. (1981). Nonparametric Standard Errors and Confidence Intervals. *Canadian Journal of Statistics*, 9, 139-172.

Ferguson, G.R., Ponikowski, C., and Coleman, J. (2010). Evaluating Sample Design Issues in the National Compensation Survey, *Proceedings of the Section on Survey Research Methods*, Alexandria, VA: American Statistical Association.

Gross, S. (1980). Median estimation in sample surveys. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 181-184.

Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, 47, 663-685.

Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of American Statistical Association*, 77, 89-96.

Johnson, A.E. (1995). Chapter 13: Business Surveys as a Network Sample. In *Business Survey Methods*, ed B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott. New York: Wiley.

Lee, H.J., Li, T., Fergusson, G.R., Ponikowski, C.H., and Rhein, B. (2012). Sample allocation for the redesigned National Compensation Survey. In *Proceedings of the Section on Survey Research Methods*, [CD-ROM]. Alexandria, VA: American Statistical Association.

Lee, H.J., and Li, T. (July, 2013). Simulation Study to Validate NCS Sample Allocation: Final Report, submitted to Bureau of Labor Statistics. Rockville, MD: Westat.

Pinciaro, S.J. (1978). An algorithm for calculating joint inclusion probabilities under PPS systematic In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 181-184.

Raghunathan, T.E, Reiter, J.P., and Rubin, D.B. (2003). Multiple Imputation for Statistical Disclosure Limitation. *Journal of Official Statistics*, 19, 1-16.

Rubin, D.B., and Schenker, N. (1986). Multiple imputation for interval estimation for simple random samples with ignorable nonresponse. *Journal of American Statistical Association*, 81, 366-374.

Sen, A.R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of Indian Society of Agricultural Statistics*, 5, 119-127.

Yates, F., and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, Series B, 109, 12-30.