Sample Allocation for the Redesigned National Compensation Survey October 2012

Hyunshik James Lee[1], Tiandong Li[1], Gwyn R. Ferguson[2],
Chester H. Ponikowski[2], Bradley Rhein[2]
[1]Westat, 1600 Research Boulevard, Rockville, MD 20850
[2]Bureau of Labor Statistics, 2 Massachusetts Avenue, NE Washington, DC 20212

Abstract:
The National Compensation Survey (NCS) conducted by the Bureau of Labor Statistics (BLS) is an establishment survey for which the sample selection methodology has been redesigned to reflect a change in the scope of the survey. The new sample design will be implemented in 2012. The main feature of the redesign is the change from a three-stage area based design to a new two-stage design. Under the new design, establishments are selected by the probability proportional to size (PPS) sampling method as the primary sampling unit (PSU) at the first-stage, and occupations again by the PPS method at the second-stage from the sampled establishments. For this new design, sample allocation needs to be done to meet the precision objectives of the survey. For this purpose, variance components are estimated, response rates are projected, and based on this information the national private industry sample is allocated to sample allocation cells defined by 23 detailed industries. This paper discusses these steps and the final sample allocation results.

Key Words: Logistic regression, viability and usability rates, between- and within-PSU variances, intra-class correlation

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, 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 redesigned NCS sample is selected using a two stage stratified design with probability proportionate to employment size (PPS) sampling at each stage. The first stage of sample selection is a probability sample of establishments in 24 pre-determined geographic area strata and the second stage is a probability selection of occupations (PSO) within the establishments. 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 areas. 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 estimate methodology is given in Chapter 8 of BLS Handbook of Methods.

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 estimates. 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 this paper estimates of variance components, design effect, and response rates are projected and used to derive sample allocations for the private industry strata in the new design. Section 2 describes the methodology used to estimate response rates and presents projected response rates. Section 3 presents the derivation of design effect formula, estimation of variance components and intra-class correlation, and estimation of the design effect. Section 4 presents the size determination formula, estimation of components of the formula, and sample allocation results. Finally, section 5 provides a summary of our findings and presents areas of possible future research.

2. Estimation of the Response Rate

The eligibility rate of NCS, also known as the viability rate, is defined as the percent of eligible establishments for this survey over all establishments in the data collection frame. The response rate, also known as the usability rate, is the percent of responding establishments over eligible ones for this survey. The ineligible establishments include three types: 1) there are no matching occupations and collected in-scope employment is zero; 2) the establishment has ceased all productive operations; and 3) the establishment is in an industry or area outside of the survey's coverage. For a responding establishment, the schedule contains at least one "USE" occupation, which is classified as such if the following are present: occupational characteristics (full-time/part time, unionized/non-unionized, and time/incentive), work schedule, and wage data.

We predicted the eligibility and response rates in separate models using a logistic regression model with predictor variables available in the sampling frame, which was developed from the current NCS sample. The predictor variables include categorical variables for the 23 major industries and the 24 NCS geographic areas, log-transformed employment size, and indicator variable of whether the employment size greater than 1 or not. The basic model without interaction terms was first fitted within each of the 45 domains (model groups) defined by crossing 5 major industry groups and 9 NCS geographic divisions (Census Divisions with adjusted boundaries to accommodate the NCS largest metropolitan areas). No interaction term was used assuming that main interaction terms would be removed by estimating the model separately for each of the 45 domains. In addition, alternative models with all the 2-way and 3-way interaction terms were tested and evaluated. The predicted response rates from the alternative models for some major industries were inconsistent to NCS experience, which may be due to model overfitting. Thus, we chose to use the basic model for sample size allocation. The fitted model was then applied to individual establishments in the frame to predict the eligibility (or response) propensities. Then the predicted propensities for the individual establishments in the frame are aggregated to calculate the sample size inflation factors by the sampling stratum.

Table 2.1 presents the combined rates (the product of the eligibility rate and the response rate) for the basic model without interaction terms, by major industries. The range of the combined rates is between 42.3% and 77.2%.

Table 2.1: The Combined Rates (the product of the eligibility rate and response rate) for the Basic Model by Major Industries

		Basic Model
Major Group	*Major Industry*	*Without interaction*
1	Mining	53.0%
	Construction	42.3%
	Manufacturing	67.2%
2	Finance (excluding Insurance)	49.1%
	Insurance Carriers and Related Activities	57.1%
	Real Estate and Rental and Leasing	45.2%
3	Education	44.9%
	Elementary & Secondary Education	74.9%
	Colleges & Universities	75.8%
4	Health and Social Assistance	63.7%
	Hospitals	77.2%
	Nursing Homes	71.0%
5	Utilities	67.1%
	Wholesale Trade	52.0%
	Retail Trade	60.3%
	Transportation and Warehousing	55.5%
	Information	49.6%
	Professional, Scientific, and Tech Services	50.5%
	Management of Companies and Enterprises	58.7%
	Admin and Support, Waste Management	44.1%
	Arts, Entertainment, and Recreation	51.5%
	Accommodation and Food Services	63.5%
	Other services except public administration	52.8%
Overall		58.0%

3. Estimation of the Design Effect

The sample allocation task was geared to achieve a certain level of precision for the Employer Costs of Employee Compensation (ECEC) series that measures the average (total) cost to employers for wages, salaries, and benefits, per employee hour worked. It is estimated from the cost data collected from sample occupations (quotes) selected from sample establishments. For a particular domain *D*, it is formally defined by:

$$\hat{R}_D = \frac{\sum_{q \in S_D} w_q y_q}{\sum_{q \in S_D} w_q x_q} \tag{3.1}$$

where S_D is the intersected part of the sample S in the domain, w_q is the final sampling weight for quote *q* that reflects both stages of sampling (i.e., the first-stage sampling of establishments and the second-stage sampling of quotes) and various adjustments (e.g., nonresponse adjustment, benchmarking, etc.), y_q is the hourly total compensation for quote *q*, and x_q is total number of employees to which the quote is referenced. A quote is "a sampled job that has been matched with an SOC (Standard Occupational Classification) occupation" and refers to a within-establishment entity (see Chapter 8 of BLS Handbook of Methods). Therefore, S_D is a collection of all quotes appearing in the sample of establishments that satisfy the definition of the domain. So, the summation in $q \in S_D$ includes all the levels of sampling stages to collect establishments that contribute the y-values. We are particularly interested in domains of major industries. It is important to distinguish w_q, which can be considered as the overall quote sampling weight (i.e., the inverse of the overall quote selection probability), from the traditional quote weight that includes x_q in the definition.

The population parameter of interest in the ECEC series is the ratio defined by:

$$R_D = \frac{\sum_{q \in U_D} y_q}{\sum_{q \in U_D} x_q} = \frac{Y_D}{X_D} \tag{3.2}$$

where U_D is the index set of quotes for the population domain *D*, and X_D and Y_D are the population domain totals of x- and y-values. Using the theory of the Taylor linearization variance estimator, an approximate variance estimator of the ratio estimator in (1) is given by:

$$\begin{aligned} \hat{V}(\hat{R}_D) &= \hat{V}\left\{\sum_{q \in S_D} w_q (y_q - \hat{R}_D x_q) \big/ \hat{X}_D\right\} \\ &= \hat{V}\left\{\sum_{q \in S_D} w_q z_q\right\} \end{aligned} \tag{3.3}$$

where $z_q = (y_q - \hat{R}_D x_q) \big/ \hat{X}_D$ is the linearized value, which is sometimes called a pseudo value, and $\hat{X}_D = \sum_{q \in S_D} w_q x_q$.

Since 23 major industries are our primary domains for sample allocation, we confine our discussion for them. So, we define D as a particular major industry i, and within this industry, establishment are indexed by j and quotes within establishment j is indexed by k. For ease of notation, we drop the domain index for the time being. We assume the following model for y_{jk}, the total cost for quote jk.

$$\begin{aligned} &y_{jk} = Rx_{jk} + e_{jk} \\ &V(y_{jk} \mid x_{jk}) = x_{jk}^2 \sigma_e^2 \\ &C(y_{jk}, y_{j'k'} \mid x_{jk}, x_{j'k'}) = \begin{cases} \rho x_{jk} x_{j'k'} \sigma_e^2 & \text{if } j = j' \text{ and } k \neq k' \\ 0 & \text{if } j \neq j' \end{cases} \end{aligned} \tag{3.4}$$

where x_{jk} is the number of employees for quote jk, and ρ is the intra-class correlation. The first part is the usual ratio model considered appropriate for the NCS data. If we take $x_{jk} = 1$ for all jk, the model in (3.4) reduces the model Gabler, Häder, and Lahiri (1999) used.

Let N be the establishment population size for the major industry and M_j be the total number of quotes in establishment j. Then the total number of quotes in the major industry is given by $M = \sum_{j=1}^{N} M_j$. Under the new design, the establishment selection probability is $p_j = nX_j / X$ and within-establishment quote selection probability is $p_{k|j} = b_j x_{jk} / X_j$, where n is the establishment sample size, $X_j = \sum_{k=1}^{M_j} x_{jk}$, and b_j the quote sample size. Then the overall probability is given by $p_{jk} = p_j p_{k|j} = nb_j x_{jk} / X$ and the overall weight is given by $w_{jk} = w_j w_{k|j} = X / nb_j x_{jk}$, where $X = \sum_{j=1}^{N} \sum_{k=1}^{M_j} x_{jk}$, which is the total population number of employees in the major industry, $w_j = p_j^{-1}$, and $w_{k|j} = p_{k|j}^{-1}$.

If R and X are known, then an approximate variance under the two-stage PPS design is given by

$$\begin{aligned} V_{PPS}(\hat{R}) &\cong V_{pps}\left(\sum_{j=1}^{n} \sum_{k=1}^{b_j} w_{jk} z_{jk} \right) \\ &= V_{pps}\left(\sum_{j=1}^{n} \sum_{k=1}^{b_j} w_{jk} \varepsilon_{jk} \right) \end{aligned} \tag{3.5}$$

where,

$$\varepsilon_{jk} \equiv z_{jk} = (y_{jk} - Rx_{jk}) / X = e_{jk} / X. \tag{3.6}$$

Here the z-value is defined slightly differently from what was defined before, where $\hat{R}$ was used before instead of R.

Under the model given in (3.4), we have

$$V(\varepsilon_{jk}) = x_{jk}^2 \sigma_e^2 \big/ X^2 \tag{3.7}$$

and

$$\begin{aligned}
&V_{PPS}\left(\sum\nolimits_{j=1}^{n}\sum\nolimits_{k=1}^{b_j} w_{jk}\varepsilon_{jk}\right) \\
&= \sum\nolimits_{j=1}^{n}\sum\nolimits_{k=1}^{b_j} w_{jk}^2 V(\varepsilon_{jk}) + \sum\nolimits_{j=1}^{n}\sum\nolimits_{k\neq k'}^{b_j} w_{jk}w_{jk'}C(\varepsilon_{jk},\varepsilon_{jk'}) \\
&= \sum\nolimits_{j=1}^{n}\sum\nolimits_{k=1}^{b_j} w_{jk}^2 x_{jk}^2 \sigma_e^2 \big/ X^2 + \sum\nolimits_{j=1}^{n}\sum\nolimits_{k\neq k'}^{b_j} w_{jk}w_{jk'}x_{jk}x_{jk'}\rho\sigma_e^2 \big/ X^2 \\
&= \sum\nolimits_{j=1}^{n}\sum\nolimits_{k=1}^{b_j} \sigma_e^2 \big/ (n^2 b_j^2) + \sum\nolimits_{j=1}^{n}\sum\nolimits_{k\neq k'}^{b_j} \rho\sigma_e^2 \big/ (n^2 b_j^2) \\
&= \frac{\sigma_e^2}{n^2}\left[\sum\nolimits_{j=1}^{n} b_j^{-1} + \rho\sum\nolimits_{j=1}^{n}(1-b_j^{-1})\right].
\end{aligned} \tag{3.8}$$

In the above derivation, we tacitly assume that there are no sample observations selected with certainty. Therefore, it is somewhat biased to the extent that the assumption is violated. Nevertheless, we believe that it still provides a useful approximation.

To derive the design effect formula, we need the variance under a simple random sample (SRS) of quotes with the same sample size as for the two-stage design. The model for the SRS situation is the same as for the two-stage (cluster) design but because there is no clustering, the intra-class correlation is zero, that is, $\rho = 0$. The total quote sample size is given by

$$m = \sum\nolimits_{j=1}^{n} b_j . \tag{3.9}$$

Under SRS, the R is estimated by $\hat{R} = \hat{Y} \big/ \hat{X}$ but now $\hat{X}$ and $\hat{Y}$ are estimates of the totals under SRS. Then the approximate variance of this estimate using the Taylor linearization method is based on the z-values given in (8). Evaluating the approximate variance under the model given in (3.4) with $\rho = 0$, we get

$$\begin{aligned}
V_{SRS}(\hat{R}) &= V_{SRS}\left(\frac{M}{m}\sum\nolimits_{j=1}^{n}\sum\nolimits_{k=1}^{b_j}\varepsilon_{jk}\right) \\
&= \left(\frac{M}{m}\right)^2 \sum\nolimits_{j=1}^{n}\sum\nolimits_{k=1}^{b_j} V(\varepsilon_{jk}) \\
&= \left(\frac{M}{m}\right)^2 \sum\nolimits_{j=1}^{n}\sum\nolimits_{k=1}^{b_j} x_{jk}^2\sigma_e^2 \big/ X^2 \\
&= \frac{1}{(m\bar{X})^2}\sum\nolimits_{j=1}^{n}\sum\nolimits_{k=1}^{b_j} x_{jk}^2\sigma_e^2 .
\end{aligned} \tag{3.10}$$

where $\bar{X} = X/M$ is the population mean of x-values.

Then the design effect for $\hat{R}$ is given by

$$
\begin{aligned}
\mathrm{Deff}(\hat{R}) &= V_{PPS}(\hat{R}) \big/ V_{SRS}(\hat{R}) \\
&= \frac{\dfrac{\sigma_e^2}{n^2}\left[\sum_{j=1}^{n} b_j^{-1} + \rho \sum_{j=1}^{n} (1-b_j^{-1})\right]}{(m\bar{X})^{-2} \sum_{j=1}^{n} \sum_{k=1}^{b_j} x_{jk}^2 \sigma_e^2} \\
&= \frac{(m\bar{X})^2 \left[\sum_{j=1}^{n} b_j^{-1} + \rho \sum_{j=1}^{n} (1-b_j^{-1})\right]}{n^2 \sum_{j=1}^{n} \sum_{k=1}^{b_j} x_{jk}^2} \\
&= \frac{m\bar{X}^2}{n^2 a}\left[\sum_{j=1}^{n} b_j^{-1} + \rho \sum_{j=1}^{n} (1-b_j^{-1})\right] \\
&= \frac{\bar{b}\bar{X}^2}{a}\left[d + \rho(1-d)\right]
\end{aligned}
\tag{3.11}
$$

where $a = \sum_{j=1}^{n} \sum_{k=1}^{b_j} x_{jk}^2 \big/ m$, which is the simple average of x_{jk}^2 's in the SRS sample, $\bar{b} = m/n$, which is the average number of quotes in the sample, and $d = \sum_{j=1}^{n} b_j^{-1} \big/ n$. Note that the observed values in the denominator are from the SRS sample, which is different from those in the numerator, although we use similar notations. If we replace the b_j 's in (3.11) by their average $\bar{b} = m/n$, then we get an interesting approximate expression;

$$
\mathrm{Deff}(\hat{R}) \cong \frac{\bar{X}^2}{a}\left[1 + \rho(\bar{b} - 1)\right] \tag{3.12}
$$

The second part in the bracket of (3.12) is exactly the same as in Kish's formula (Kish, 1987). Using the inequality between the arithmetic mean and the harmonic mean of b_j 's, we can show that (3.12) is smaller than (3.11). If we use $A = E(a) = \sum_{j=1}^{N} \sum_{k=1}^{M_j} x_{jk}^2 \big/ M$, which is the expected value of *a* instead of a in (3.12), we have

$$
\mathrm{Deff}(\hat{R}) \cong \frac{\bar{X}^2}{A}\left[1 + \rho(\bar{b} - 1)\right] \tag{3.13}
$$

Using the Cauchy-Schwarz inequality, we can show that $\bar{X}^2 \big/ A \leq 1$. Therefore, the design effect can be less than 1 if the first factor is much smaller than 1.

We need to estimate this design effect to be used for sample allocation. For the numerator, we can plug in a sample estimate from the current sample for unknown quantities (i.e., $\bar{X}^2$ and ρ). However, for the denominator, since *a* is based on the SRS sample, we cannot simply use the current sample in calculating *a*. Nevertheless, it would be reasonable to use the current sample estimate for *A*, the expected value of *a*. A good sample estimate of this using the current sample is

$$
\hat{A} = \sum_{j=1}^{n} \sum_{k=1}^{b_j} w_{jk} x_{jk}^2 \big/ \hat{M} \tag{3.14}
$$

where $\hat{M} = \sum_{j=1}^{n} \sum_{k=1}^{b_i} w_{jk}$.

Then an estimated Deff for the domain is given by

$$\hat{\delta} = \frac{\bar{b}\hat{\bar{X}}^2}{\hat{A}}\left[d + \hat{\rho}(1-d)\right] \tag{3.15}$$

We propose to use the current sample $\bar{b}$ and *d* since these are not population means.

Now we need an estimate of the intra-class correlation. The current design is a three-stage design, but if we estimate variance components for the three stage sampling units (i.e., areas in the first-stage, establishments in the second-stage, and quotes in the third stage), we can estimate ρ using these variance components.

Denoting each stage variance component as $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$, and $\hat{\sigma}_3^2$, we can estimate ρ that will be used in (3.15) for the new design with two stages of sampling by:

$$\hat{\rho} = \frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \hat{\sigma}_3^2} \tag{3.16}$$

Another approach is to estimate $\hat{\sigma}_1^2 + \hat{\sigma}_2^2$ as one component by using establishments as PSUs, and this is a better approach because we do not need two separate components and we will have one less parameter to estimate, hence, larger degrees of freedom. Using the SAS procedure, we get 23 estimated intra-class correlations, $\hat{\rho}_i$, $i = 1,2,3,...,23$.

The design effect estimate given in (3.15) needs further estimates for the unknown population quantities to be obtained from the current sample. We estimate *X* and *A* by the usual estimators for a particular major industry *i* using the current sample as follows:

$$\begin{aligned} \hat{X}_i &= \sum\nolimits_{j=1}^{n_i} \sum\nolimits_{k=1}^{b_{ij}} w_{ijk} x_{ijk} \\ \hat{A}_i &= \sum\nolimits_{j=1}^{n_i} \sum\nolimits_{k=1}^{b_{ij}} w_{ijk} x_{ijk}^2 \Big/ \hat{M}_i \\ \hat{M}_i &= \sum\nolimits_{j=1}^{n_i} \sum\nolimits_{k=1}^{b_{ij}} w_{ijk}. \end{aligned} \tag{3.17}$$

The weight in (3.17) is the quote sampling weight of the current sample. Plugging these values in (3.15), we obtain a design effect estimate $\hat{\delta}_i$ for the new two-stage design. The n_i and m_i in (3.15) are supposed to be sample sizes for the new design. However, since we are estimating the design effect using the current data, we have to use the realized sample sizes for the current data. We believe that slightly different sample sizes will not affect the estimated design effect much. We use the original quote sample size at the time of sample selection for b_{ij}'s, which is predetermined according to the size of the establishment (see table 4.1).

4. Sample Allocation for the NCS New Design

The formula (3.15) gives an estimate of the design effect but it was originally defined as the ratio of the variance of an estimate for the population parameter obtained from a design to the variance of the sample estimate for the parameter under simple random sampling. Hence, ignoring the finite population correction (fpc), we can write

$$\delta_i = \frac{V(\hat{Z}_i)}{M_i^2 S_i^2 / m_i} \tag{4.1}$$

where S_i^2 is the population variance of the *z*-values, M_i is the population size of quotes, and m_i is the quote sample size under the new design for major industry *i*. Note that the denominator in (4.1) is the variance of the simple random sample estimate for the population total Z_i . Then the variance of $\hat{R}_i$ can be given by

$$V(\hat{R}_i) \cong V(\hat{Z}_i) = \frac{\delta_i M_i^2 S_i^2}{m_i} \tag{4.2}$$

Note that the sample size m_i is in terms of the number of quotes, not of establishments.

The initial precision requirement says that the square root of the predicted variance given in (4.2) should be 1 percent of the population ratio, R_i or less for as many major industries as possible. We first try to achieve the precision goal for every major industry, and then the precision requirement can be written as

$$V(\hat{R}_i) \cong \frac{\delta_i M_i^2 S_i^2}{m_i} = \left(0.01 R_i\right)^2 \tag{4.3}$$

Or

$$m_i = \frac{\delta_i M_i^2 S_i^2}{\left(0.01 R_i\right)^2} \tag{4.4}$$

To calculate this sample size, we need to use an estimate for the unknown population quantities, δ_i, S_i^2, and R_i . Using the current sample data, δ_i will be estimated by (3.15), M_i by $\hat{M}_i = \sum_{j=1}^{n} \sum_{k=1}^{m_i} w_{ijk}$, R_i by (3.1), and S_i^2 by the following:

$$\hat{S}_i^2 = \frac{\sum_{j=1}^{n} \sum_{k=1}^{m_i} w_{ijk} (z_{ijk} - \bar{z}_{ij})^2}{\hat{M}_i - 1} \tag{4.5}$$

where $\bar{z}_{ij} = \sum_{j=1}^{n} \sum_{k=1}^{m_i} w_{ijk} z_{ijk} / \hat{M}_i$. One important aspect of estimating $\bar{b}_i$ and d_i is that they are estimated by using originally selected quote sample size b_{ij} but using the current establishment sample size n_i' as shown below:

$$\bar{b}_i = \sum_{j=1}^{n_i'} b_{ij} / n_i'$$
$$d_i = \sum_{j=1}^{n_i} b_{ij}^{-1} / n_i' \tag{4.6}$$

The b_{ij} are selected according to the schedule given in Table 4.1.

Finally the allocated sample size for major industry *i* for the new design will be determined by

$$m_i = \frac{\hat{\delta}_i \hat{M}_i^2 \hat{S}_i^2}{\left(0.01\hat{R}_i\right)^2} \tag{4.7}$$

Table 4.1: The Selection Schedule of Quotes

Employment Size	*Number of Quotes to Be Selected*
1	1
2	2
3	3
4 – 49	4
50-249	6
250+	8

The sample size for the new design in terms of the number of establishments can then be obtained by

$$n_i = \frac{m_i}{\bar{b}_i} \tag{4.8}$$

As $n = \sum_{i=1}^{23} n_i$ exceeds far more than the fixed total sample size (9,754 establishments after inflation of the combined response rate) for the new design, we adjusted n_i to comply the fixed total sample size constraint. This adjusted sample size was then inflated by multiplying the inverses of the eligibility rate and response rate estimated in section 2.

Table 4.2 presents estimated components in formulas (4.7) and (4.8) and the estimated sample sizes for 23 major industries with a precision goal of RSE = 1%. The "Total Establishment Sample Size" is the number of establishments after the adjustment of the eligibility rate and response rate estimated the basic model presented in Table 2.1. To achieve this precision level for each industry, a total number of 56,161 establishments is needed, which is much larger than the fixed sample size for the new design. The sample size was adjusted to achieve the fixed sample size of 9,754 establishments after inflation of the response rate obtained from the basic model.

Table 4.3 presents the inflated sample size for sample selection and expected relative standard error (RSE) by major industry that sums to the fixed total sample size of 9,754 establishments. With this sample size, the expected RSE under new design increased to

2.40 percent for each major industry. It is not shown but the overall national RSE is expected to be 0.91 percent, which is less than 1 percent.

Table 4.2: Sample Selection Size for 23 Major Industries with a Precision Goal of RSE = 1%

Major Group	*Industry*	*Average Quote Sample Size*	*DEFF*	*Total # of Quote Resp.*	*Total # of Est Resp.*	*Total Est Sample Size*	*Expecte d RSE (%) under New Design*
1	Mining	5.488	0.368	6,326	1,153	2,175	1
	Construction	4.700	0.224	6,628	1,410	3,334	1
	Manufacturing	6.678	0.072	7,560	1,132	1,685	1
2	Finance (excluding Insurance)	5.693	0.133	24,089	4,231	8,618	1
	Insurance Carriers and Related Activities	5.958	0.135	2,242	376	659	1
	Real Estate and Rental and Leasing	4.584	0.259	3,159	689	1,525	1
3	Education (rest of)	5.000	0.110	2,539	508	1,131	1
	Elementary & Secondary Education	5.391	0.319	3,714	689	920	1
	Colleges & Universities	7.526	0.090	3,722	495	652	1
4	Health and Social Assistance (rest of)	5.363	0.054	7,118	1,327	2,084	1
	Hospitals	7.878	0.067	2,737	347	450	1
	Nursing Homes	6.097	0.180	1,791	294	414	1
5	Utilities	6.281	0.226	1,684	268	400	1
	Wholesale Trade	5.182	0.250	5,683	1,097	2,109	1
	Retail Trade (rest of)	5.411	0.175	1,967	364	603	1
	Transportation and Warehousing	6.662	0.040	38,162	5,728	10,321	1
	Information (rest of)	6.459	0.084	2,522	390	787	1
	Professional, Scientific, and Tech Services	5.460	0.147	6,522	1,194	2,365	1
	Management of Companies and Enterprises	7.062	0.134	5,223	740	1,260	1
	Admin and Support, Waste Management	5.976	0.074	6,350	1,063	2,409	1
	Arts, Entertainment, and Recreation	6.086	0.099	2,335	384	745	1
	Accommodation and Food Services	5.416	0.179	2,490	460	724	1
	Other services except public administration	4.569	0.351	26,039	5,699	10,793	1
Total				170,600	30,037	56,161	

Table 4.3: Allocated sample selection size and expected relative standard error (RSE) by major industry, with the total sample size of 9,754 establishments

		Allocation Based on Basic Model			
Major Group	*Industry*	*Total # of Quote Resp.*	*Total # of Establish Resp.*	*Total Establish Sample Size*	*Expected RSE(%) under New Design*
1	Mining	1,099	200	378	2.40
	Construction	1,151	245	579	2.40
	Manufacturing	1,313	197	293	2.40
2	Finance (excluding Insurance)	4,184	735	1,497	2.40
	Insurance Carriers and Related Activities	389	65	114	2.40
	Real Estate and Rental and Leasing	549	120	265	2.40
3	Education (rest of)	441	88	196	2.40
	Elementary & Secondary Education	645	120	160	2.40
	Colleges & Universities	646	86	113	2.40
4	Health and Social Assistance (rest of)	1,236	231	362	2.40
	Hospitals	475	60	78	2.40
	Nursing Homes	311	51	72	2.40
5	Utilities	293	47	69	2.40
	Wholesale Trade	987	190	366	2.40
	Retail Trade (rest of)	342	63	105	2.40
	Transportation and Warehousing	6,628	995	1,792	2.40
	Information (rest of)	438	68	137	2.40
	Professional, Scientific, and Tech Services	1,133	207	411	2.40
	Management of Companies and Enterprises	907	128	219	2.40
	Admin and Support, Waste Management	1,103	185	418	2.40
	Arts, Entertainment, and Recreation	406	67	129	2.40
	Accommodation and Food Services	433	80	126	2.40
	Other services except public administration	4,522	990	1,875	2.40
Total		29,630	5,217	9,754	

5. Conclusions

There are two main components in this sample allocation task for the new NCS design. The first is the estimation of combined response rate that is the product of the eligibility and response rates. This was accomplished using the logistic regression model that relates the response propensity with predictor variables available in the frame file. We first used the basic model separately estimated for each of 45 model groups. Realizing that some interaction terms are important, we also used an alternative model that includes all possible two- and three-way interaction terms. The estimated response rates by the basic model are generally larger than those estimated by the alternative model. So, we also took the average of the two rates. These rates are used to calculate the sample size needed to select the field sample.

To determine the required sample size that would be needed to achieve a certain level of precision requirement for the ECEC series, we developed a design effect formula that is suitable for the NCS design and estimated the design effect for the new design. The estimated design effect was then used to calculate the required sample size for each of the 23 major industries. This sample size was way too big, so it was substantially reduced so that the sample size adjusted for the combined response rate to obtain the appropriate sample size for field work be summed up to the overall sample size of 9,754 BLS determined. The expected RSE for this final allocation 2.40 percent with the response rate by the basic model, and the overall national level RSE is slightly over 0.9 percent.

We made many assumptions to derive the design effect formula that is central to this sample allocation task. There is no doubt that those assumptions are not perfect and deviate from the reality to a certain extent, and the derived formula is somewhat invalid to that extent. Nevertheless, the results look reasonable in comparison with historical information of the survey.

As mentioned earlier, the eligibility and response rates estimated by the basic model are generally higher than those estimated by the alternative model. We do not know the reason, but it would be interesting to investigate and find an explanation.

References

Ferguson, Gwyn R., Coleman, Joan, Ponikowski, Chester H. (2011), "Update on the Evaluation of Sample Design Issues in the National Compensation Survey", *2011 Proceedings of the Section on Survey Research Methods*, Alexandria, VA: American Statistical Association. http://www.bls.gov/osmr/abstract/st/st110230.htm.

Ferguson, Gwyn R., Ponikowski, Chester, and Coleman, Joan (2010), "Evaluating Sample Design Issues in the National Compensation Survey", *2010 Proceedings of the Section on Survey Research Methods*, Alexandria, VA: American Statistical Association, http://www.bls.gov/osmr/abstract/st/st100220.htm.

Gabler, S., Häder, S., and Lahiri, P. (1999). A model based justification of Kish's formula for design effect for weighting and clustering. *Survey Methodology*, 25, 105-106.

Kish, L. (1987). Weighting in Deft. *The Survey Statistician*, June 1987.

Kish, L. (1992). Weighting for unequal p_i. *Journal of Official Statistics*, 8, 183-200.

U.S. Bureau of Labor Statistics (2008) *BLS Handbook of Methods*, National Compensation Measures, Chapter 8. http://www.bls.gov/opub/hom/pdf/homch8.pdf

U.S. Bureau of Labor Statistics (2008) *BLS Handbook of Methods*, Occupational Employment Statistics, Chapter 3. http://www.bls.gov/opub/hom/pdf/homch3.pdf