Evaluating the Effect of Dependent Sampling on National Compensation Survey Earnings Estimates

Omolola E. Ojo[1], Chester Ponikowski[1]
[1]U.S. Bureau of Labor Statistics, 2 Massachusetts Ave., NE, Room 3160, Washington, DC 20212

Abstract
The National Compensation Survey is an establishment survey that provides measures of occupational earnings and benefits. The private industry establishment sample is divided into five sample rotation panels, with the panels being fully replaced over a five-year period. Data are updated periodically for each selected establishment and occupation until the panel in which the establishment was selected is replaced. The current sampling method allows for establishments to appear in more than one rotation panel. This can increase the reporting burden if a new set of occupations is selected for an overlapping establishment. This research uses sample simulations to explore an alternative design in which establishments selected in a sample panel would not be eligible for selection in subsequent panels of the rotation and discusses potential pitfalls in employing such a sampling approach.

Key Words: Multi-stage sampling, sampling cells, respondent burden

1. Introduction

Data collected by the National Compensation Survey (NCS) provides employee wage estimates for areas and industries across the fifty United States and the District of Columbia. The private industry establishment sample for the survey is distributed among five rotating panels of independent samples selected annually under probability proportional to employment size (PPS) sampling.

With independent sampling, each of the rotating panels are selected from the full frame of the target population regardless of what units already exist in the sample, allowing for establishments to be in multiple panels of the sample. Establishments selected with certainty – very large, self representing establishments – are identified once and remain fixed for five years. These certainty units are collected in every panel of the rotation. Along with overlapping non-certainty establishments, multi-year certainty establishments present a respondent burden issue. As an alternate design, dependent sampling offers an avenue to alleviating respondent burden by eliminating the possibility of overlapping establishments between rotating panels. This paper will provide an overview of the NCS, share the details of our simulation study for measuring the effect of dependent sampling, and present the comparative results of independent versus dependent sampling with regards to the NCS earnings estimates.

2. The National Compensation Survey

The NCS is an establishment survey of occupational wages and benefits conducted by the Bureau of Labor Statistics (BLS). Data from the survey are used to produce three general products: employment cost data, employee benefits data, and national and locality wage data. The employment cost data include the Employment Cost Index (ECI) and Employer Costs for Employee Compensation (ECEC). The ECI provides a series of indexes that track quarterly and annual changes in wages and benefit costs. The ECEC provides quarterly cost level information on the cost per hour worked for pay and benefits. Employee benefits data include incidence and provisions of selected employee benefit plans and are published once a year. National and locality wage data include annual publications of occupational wages for a sample of the NCS locality areas, census divisions, and for the nation as a whole. The scope of the NCS covers all state and local governments and private sector industries other than agriculture. Private household businesses are excluded from the survey.

The BLS Quarterly Census of Employment and Wages (QCEW) serves as the sampling frame for the NCS and supplied the administrative data for this study. The QCEW is populated from individual State Unemployment Insurance (UI) files of establishments – a mandatory filing for all established businesses.

The integrated NCS sample currently consists of five rotating replacement sample panels. Each of the five sample panels are 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, sampling with probability proportional to employment size (PPS) at each stage. The first stage 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.

The selection of Primary Sampling Units (PSUs) occurs at the national level across geographic areas. They are based on the 2003 Office of Management and Budget (OMB) area definitions. Areas are redefined every 10 years based on the U.S. decennial census. Under the OMB definition there are three types of statistical areas - Metropolitan, Micropolitan, and Combined Statistical Areas. Combined Statistical Areas (CSAs) are defined as a combination of adjacent Metropolitan and Micropolitan areas that meet specified conditions. Outside of these areas exists a number of counties. These counties are referred to as outside Core Based Statistical Areas (CBSA). For selection purposes these outside CBSA's are organized into clusters to create more heterogeneous primary sampling units.

In 2004, a new area sample was selected for the NCS. This sample contained 152 areas. In this sample 57 areas were selected with certainty, which are defined as areas having employment greater than 80 percent of the final sampling interval. The remaining areas consisted of 60 non-certainty metropolitan areas, 22 non-certainty Micropolitan areas, and 13 non-certainty outside CBSA county clusters.

The second stage of this design occurs at the establishment level within each selected area. In this stage, a sample of non-agricultural private business establishments and state and local government operations are selected from an area frame that is stratified by ownership and industry. Within each stratum we employ PPS systematic sampling with frame employment as the measure of size (MOS). Certainty units – units with probability

of selection greater than or equal to one – are identified and removed from the sampling frame. These establishments are automatically included in the sample with a sampling weight set to one. Furthermore, in order to assure that each establishment has a chance of selection, any establishment having a frame employment of zero is given an adjusted employment of one (1) for sampling purposes. After the sample of establishments is selected, it is used for the third stage of the sampling process.

The third stage of this design occurs at the occupational level within each selected establishment. A sample of jobs is drawn from each of these establishments using PPS systematic sampling. To ensure that jobs are defined consistently across all establishments, the Standard Occupational Classification (SOC) manual is used to classify these jobs based upon their occupational duties. After the selection and classification of jobs we create our smallest aggregate unit known as a quote, which is a distinct combination of occupation, time or incentive pay, job level, union membership, and full-time or part-time status.

This study focuses on the second stage of sampling, the establishment sample, with regards to our national and locality earnings data. By means of a series of simulations, our study examines if and how dependent sampling alters estimates of wages across 23 private industry strata of the National Compensation Survey.

3. Independent Sampling vs. Dependent Sampling

For this paper, *independent sampling* refers to the current sampling methodology of the NCS in which each of the five annual sample rotation panels that make up a full NCS sample are sampled independently. As previously noted, this sampling methodology allows for overlapping units between rotating sample panels. The key disadvantage of such a methodology is that frequent requests of data from these multi-panel establishments pose a potential burden to our respondents -- and may consequently thwart their desire to respond.

Dependent sampling refers to an alternate sampling method which would allow for units to appear in only one of the rotating panels of the sample. After selection of the first panel of the sample, each panel thereafter is dependent on what units already exist in the sample. Units currently in the sample are excluded from subsequent selection frames and therefore cannot be selected again for the sample for the remainder of the rotation.

Dependent sampling could ease respondent burden. Easing respondent burden may positively affect an establishment's tendency to respond. As noted by Duncan and Kalton (1987), a useful method for minimizing burden placed on responding units for a rotating panel survey, such as the NCS, may be for the survey design to condition their sampling methodology to assure that units are not included in more than one of the panels. However, dependent sampling could potentially impact the quality of NCS wage estimates due to variable sampling weights that result from the exclusion of establishments in subsequent sampling frames.

4. Empirical Study

Using the BLS Quarterly Census of Employment and Wages (QCEW) as the sampling frame, we simulated 100 full samples across all 152 NCS areas for four categories of

sampling schemes – independent sampling under a five-year rotation, independent sampling under a three-year rotation, dependent sampling under a five-year rotation, and dependent sampling under a three-year rotation. The three-year sample rotation was studied as it is a proposed option for a potential redesign of the NCS. The simulated samples followed the most current NCS sample allocation available at the beginning of the study. Based on the current frame and establishment sample allocation, the certainty units were identified and removed from the QCEW frame to create the non-certainty selection frame. The allocation for non-certainty units were divided into three or five panels, respective of the rotation scheme.

Independent samples were simulated identically to the current NCS sampling methodology – non-certainty units were drawn independently from the same frame, certainty establishments were added to every collection panel, then the panels are combined to make the full NCS sample and the overlapping units are flagged.

The same certainty units identified for the independent sample simulations are also used for the dependent samples. For dependent sampling however, the certainty units are randomly assigned to only one of the panels so that each panel contains a representative sub-sample of certainties. To simulate dependent samples, units for the first panel of the rotation are selected from the full non-certainty frame. Those selected units are then identified in the frame by a unique number and removed, creating the selection frame for the second panel. To select the third sample panel, units selected in the preceding two panels are excluded from the frame. The same pattern is followed when selecting five sample panels.

In calculating the wage estimates, we utilized administrative data from the QCEW, which includes the total wages for each establishment along with the employment count for each of the three months of the reference quarter. This wage total represents the total wages of all employees in the establishment for the three months noted. We assumed that the earnings estimates derived from QCEW data are highly correlated with average hourly earnings estimates derived from the NCS data. It is important to note that the estimates for this research are at an establishment level while the NCS reports earnings at an occupational level. In the event that a sampled establishment has a zero employment is all three months of the quarter, the unit is discounted and excluded from this wage calculation. Equation 1 illustrates how we calculated the average monthly wage per employee in each establishment. For the strata estimates in equations two and three, the employment in month 3 was used as the representative employment count. Therefore if a unit has a zero employment in month three, it is discounted from the wage calculations.

Equation 1: Formula for estimating the average wage per employee for each unit

$$\overline{x} = \frac{Q}{E1 + E2 + E3}$$

Where,

$\overline{\boldsymbol{x}}$: Average monthly wage per employee

$\boldsymbol{Q}$: Total quarterly wages for the establishment

$\boldsymbol{E(X)}$: Number of employees in the establishment in month 'X' with X = 1, 2, and 3

The average monthly wage is then used to calculate the weighted monthly wage of an employee in the sampling cell (industry strata) within the designated area. For each area, the average monthly wage for an employee in a given industry is calculated using equation 2.

Equation 2: Formula for estimating monthly wage for an employee in a given industry stratum

$$\overline{x}_s = \frac{\sum_{i=1}^{n}(\overline{x}_i)(w_i)(E_i)}{\sum_{i=1}^{n}(w_i)(E_i)}$$

Where,

$\overline{x}_s$: Sample average monthly wage per employee in industry stratum s
$\overline{x}_i$: Average monthly wage per employee in establishment *i*
w_i: Sampling weight of establishment *i.*
E_i: Number of employees in the establishment *i.*
n: Number of sampled establishments in stratum *s*

For comparative analysis, the average monthly wages were also calculated for the entire sampling frame. We calculated the average monthly wage per employee for each industry stratum in the QCEW using equation 3.

Equation 3: Calculating the monthly wage for an employee in a given industry stratum in the frame

$$\overline{X}_S = \frac{\sum_{i=1}^{N}(\overline{x}_i)(E_i)}{\sum_{i=1}^{N}(E_i)}$$

Where,

$\overline{X}_S$: Frame average monthly wage per employee in stratum *s*
$\overline{x}_i$: Average monthly wage per employee in establishment *i*
E_i: Number of employees in the establishment *i.*
N: Number of frame establishments in stratum *s*

Once all these calculations were done for each of the simulated samples as well as the full sampling frame, the wage estimates for each domain of interest were summarized as the average (arithmetic mean) of all 100 simulated samples for each of the four sampling methodologies being studied.

As a precision measurement, we then calculated the absolute percent deviation of the sample estimate from the congruent wage on frame using the formula in Equation 4.

Equation 4: Formula for calculation the absolute percent deviation of the sample estimate from the actual wage on the sampling frame for a given sample strata (Note: See equations 2 and 3 for variable meanings)

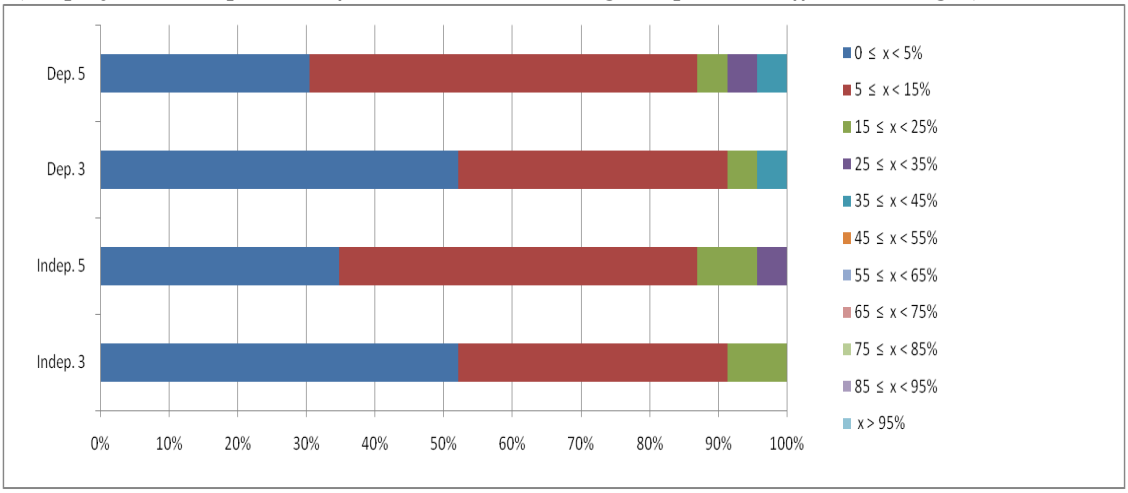
$$\left| \frac{\overline{x}_s - \overline{X}_s}{\overline{X}_s} \right| \times 100$$

5. Results and Analysis

For both the dependent and independent sampling methodologies, the three-panel rotation produces more precise sampling estimates. This is expected as each of the three panels contains more sampling units compared to the same allocation being distributed among five panels. As highlighted by Cochran (1963), the estimate of sampling variance is a known function of the sample size and we would therefore expect a higher precision in estimates with a larger sample size. This is demonstrated in Chart 1 below. Chart 1 illustrates the distribution of the total estimates among the given ranges of the percent deviation of sample wage estimates from the actual wage as noted on the frame – this is shown for the national estimates based on the full sample for each of the four sampling categories studied. Note that the national estimates are calculated using both the establishment sampling weights and the area (PSU) weights.

Chart 1: National Estimates for 23 Industry Strata - Distribution of Percent Differences of Sample Estimates from Frame Wages by Sampling Methodology

(Displayed as the percent of total estimates in the given percent difference ranges)



The highest precision in sample estimates will fall into a percent deviation of somewhere between zero and five percent (the leftmost blue shaded area in chart 1). For example under a three-panel sample rotation, approximately 52% of sample strata estimates fell into the highest precision range ($0 \leq x < 5\%$) for both independent and dependent sampling methodologies. However under a 5-panel sample rotation, approximately 36% of independent sample strata estimates fell into the highest precision range compared to a slightly lower 32% of sample strata estimates from dependent samples.

After studying the full samples, we examined the dependent sampling effect on individual panels by focusing on two individual NCS areas. We studied the New York City Combined Statistical Area (CSA) and the Miami, OK Micropolitan area in order to investigate the potential effect of sampling proportion. Currently the NCS samples about 0.21% of all establishments in the New York CSA compared to 8.72% of Miami, OK. One may anticipate that in sampling such a small proportion of a population, like NCS does for the New York CSA, dependent sampling may not reveal a notable difference in the precision of estimates when compared to independent sampling. This is because the individual panel allocations of such an area will be an even smaller proportion of the full frame, so the exclusion of sampled units from the frame is relatively insignificant. Conversely, sampling a larger proportion of a smaller Micropolitan area may yield a more noticeable difference between the sampling methodologies. Tables 1 and 2 below illustrate the precision of strata estimates, as a comparison of independent versus dependent sampling, for the New York CSA and Miami, OK area, respectively. These are shown for the full sample as well as by sampling panel for a three-panel sample rotation design.

Table 1: Comparison of the precision of independent versus dependent sample estimates by sampling panel for the **New York City Combined Statistical Area**, three-panel rotation

	INDEPENDENT Sampling				DEPENDENT Sampling			
Range of Percent Difference of Estimate (x)	Full	Panel 1	Panel 2	Panel 3	Full	Panel 1	Panel 2	Panel 3
0 ≤ x < 5%	82.6%	73.9%	86.4%	81.8%	91.3%	73.9%	77.3%	81.8%
5 ≤ x < 15%	17.4%	21.7%	9.1%	13.6%	4.3%	21.7%	13.6%	13.6%
15 ≤ x < 25%	0.0%	4.3%	4.5%	0.0%	0.0%	0.0%	9.1%	4.5%
25 ≤ x < 35%	0.0%	0.0%	0.0%	4.5%	4.3%	4.3%	0.0%	0.0%
x ≥ 35	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Table 2: Comparison of the precision of independent versus dependent sample estimates by sampling panel for the **Miami, OK Micropolitan Area**, three-panel rotation

	INDEPENDENT Sampling				DEPENDENT Sampling			
Range of Percent Difference of Estimate (x)	Full	Panel 1	Panel 2	Panel 3	Full	Panel 1	Panel 2	Panel 3
0 ≤ x < 5%	71.4%	60.0%	66.7%	90.0%	71.4%	54.5%	44.4%	11.1%
5 ≤ x < 15%	21.4%	30.0%	33.3%	0.0%	28.6%	27.3%	44.4%	55.6%
15 ≤ x < 25%	7.1%	10.0%	0.0%	0.0%	0.0%	9.1%	11.1%	22.2%
25 ≤ x < 35%	0.0%	0.0%	0.0%	0.0%	0.0%	9.1%	0.0%	11.1%
35 ≤ x < 45%	0.0%	0.0%	0.0%	10.0%	0.0%	0.0%	0.0%	0.0%
x ≥ 45	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

In the first row of table 1, which displays the percentage of strata estimates that fall into the highest precision range, it is interesting to note that the full sample estimates under dependent sampling yielded a larger proportion of estimates that fall into the highest precision range. Furthermore, the overall precision of the dependent sampling estimates seems to increase from the first panel to the third. While this appears to be counterintuitive results, it may be useful to note that sampling such a small proportion of the densely populated New York City CSA provides a greater opportunity of selecting a large number of unique, yet representative establishment samples. So we can theorize that in order to have an optimal opportunity of noticing a more dependable trend of the impact of the two sampling methodologies – keeping the same sampling proportion – one would need significantly more simulations than the 100 that we conducted.

In table 2, we notice in Miami, OK that while the proportion of strata estimates that fall into the highest precision range in the full sample are equal for both sampling methodologies, there is a different story to tell when looking in the individual panels. With independent sampling, we would expect the precision of the estimates to generally be consistent across all panels. However, table 2 shows an increase from the first panel to the third. Dealing with such a small area like Miami, OK, for which the NCS has a very small sample allocation, estimates tend to be volatile from panel to panel. As noted for the New York City CSA, conducting more simulations may exhibit a more dependable trend in the estimates. With dependent sampling, however, the precision of the estimates decreases. It is important to note that with Miami, OK being sparsely populated, coupled with the fact that PPS systematic sampling is used, the removal of establishments from previous panels may leave smaller and potentially less representative establishments to be selected in latter panels.

Dependent sampling poses a critical issue in terms of the distribution of the certainty establishments. For the simulated dependent samples, the certainty units were randomly distributed among the five or three panels. Being that the panel distribution for these self-representing units does not take into account their area and industry stratification as the non-certainty samples does, there tends to be an imbalance in the area wages among certainty establishments between the rotating panels. This is because of the potential for certainty units in higher paying industries to be more dominant (when randomly assigned) in some panels, and units of lower paying industries in others. Tables 3 and 4 illustrate this.

Table 3: Unstratified average monthly earnings in the Miami, OK Micropolitan area, simulated five-panel rotation with dependent sampling

panel	NonCertainty	Certainty	All Units
1	$2,410.23	$3,297.20	$2,786.18
2	$2,045.50	$2,033.42	$2,224.35
3	$2,564.76	$1,742.77	$2,272.82
4	$2,313.59	$2,787.56	$2,610.96
5	$2,365.50	$2,305.47	$2,392.25
Full Sample	**$2,320.12**	**$2,503.04**	**$2,439.92**

In the dependent samples simulated under a five-panel rotation for Miami, OK, as shown in table 3, the average earnings of all certainty units have a range of $1,554.43 (from

$1,742.77 in panel 3 to $3,297.20 in panel 1) between panels, while the non-certainty counterparts have a significantly smaller range of $519.26 43 (from $2,224.35 in panel 2 to $2,786.18 in panel 1).

Table 4: Unstratified average monthly earnings in the entire New York City Combined Statistical Area, simulated five-panel rotation with dependent sampling

panel	NonCertainty	Certainty	All Units
1	$5,457.87	$11,835.90	$5,760.65
2	$5,480.68	$6,175.44	$5,581.41
3	$5,684.00	$8,151.90	$5,805.99
4	$5,535.83	$9,833.06	$5,713.98
5	$5,423.82	$8,482.83	$5,556.63
Full Sample	**$5,516.69**	**$9,257.51**	**$5,602.71**

In the New York CSA, as shown in table 4, the range of average earnings of certainty units between panels is a sizeable $5,660.46 (from $6,175.44 in panel 2 to $11,835.90 in panel 1) compared to a relatively meagre range of $260.18 (from $5,556.63 in panel 5 to $5,805.99 in panel 3) between the non-certainty sample panels.

In employing a dependent sampling methodology, the design would need to incorporate a statistically sound controlling method for distributing certainty units among the panels, so to ensure more consistency between rotating panels.

6. Summary and Future Work

For the five-panel sample simulations, this initial study reveals a slightly higher precision of the national estimates produced by independent samples compared to the dependent samples. Furthermore, in smaller areas with larger sampling proportions, we observed that dependent sampling shows a slight decrease in the precision of estimates in latter panels of the rotation. To study how dependent sampling would affect other products of the National Compensation Survey, further research will include the impact on estimates of change over time. We also plan on studying benchmarking of panel weights of dependent samples, making them more representative of the full target population. Lastly, future work may also include research into other methods of distributing certainty units across dependent sample panels.

References

Cochran,W.G. (1963), Sampling Techniques, New York: John Wiley & Sons, Inc.

Duncan,G. J. And Kalton,G. (1987). Issues of Design and Analysis of Surveys Across Time. April 1987 International Statistical Review Vol. 55, No.1, The Netherlands: International Statistical Institute.

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

Izsak Y, Ernst, L. R., McNulty E,. Paben, S. P., Ponikowski, C. H., Springer G., and Tehonica, J. (2005). Update on the Redesign of the National Compensation Survey.

2005 Proceedings of the American Statistical Association, Section on Survey Research Methods, [CD-ROM], Alexandria, VA: American Statistical Association

Sharot, T. (1991). Attrition and Rotation in Panel Surveys. Journal of the Royal Statistical Society. Series D (The Statistician), Vol. 40, No. 3. London, United Kingdom: Royal statistical Society.

U.S. Bureau of Labor Statistics (2008) *BLS Handbook of Methods*, National Compensation Measures, Ch. 8. http://www.bls.gov/opub/hom/homch8_a.htm

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