Effect of Benchmarking by Industry and Size Class on National Compensation Survey Estimates

Christopher J. Guciardo
U.S. Bureau of Labor Statistics
2 Massachusetts Ave., NE, Room 3160, Washington, DC 20212

Abstract

Benchmarking is often used in establishment surveys to adjust sample weights to match the current distribution of the population of interest. In the National Compensation Survey (NCS), an establishment survey of employer compensation costs conducted by the Bureau of Labor Statistics, the sample reference period is several quarters prior to the reference period for the estimates. Hence, the weight of each sampled establishment is adjusted to match the distribution of current employment by industry from the Quarterly Census of Employment and Wages program. NCS data indicate that compensation costs also are correlated with establishment size. This research studies the effect of splitting each industry cell further by establishment size class for benchmarking purposes. Several size class partitions are being studied. First, compensation costs for the existing sample are derived using benchmarking by industry and size class, and then compared to the current estimates. Second, a multi-sample simulation is being used to assess the effect of benchmarking by industry and size class on the mean squared error of compensation estimates. Results and analysis are presented in the paper.

1. Introduction

The National Compensation Survey (NCS) is a quarterly survey of wages and benefits conducted by the Bureau of Labor Statistics (BLS). The NCS microdata support several products, including the Employer Costs for Employee Compensation (ECEC), which publishes estimates of the mean hourly costs of wages and benefits. Data are collected from a rotating sample of business establishments and state and local government entities. Each new rotation group is sampled from the most recently available frame, which is several quarters old by the time the sample is selected. For the rotation groups used in this research, the delay is three to five quarters. Each rotation group is first used in estimation about eight quarters after it is sampled. The microdata for a group are updated each quarter until the group rotates out.

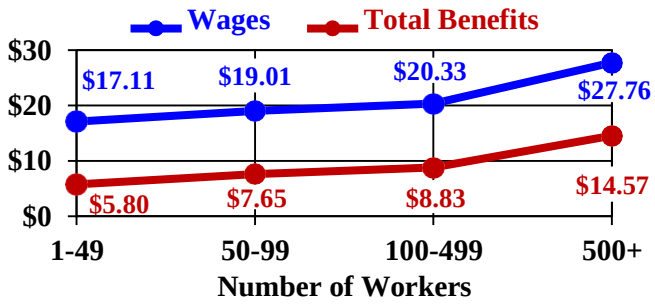
For each rotation group, the sampling process yields sample weights, whose sum is an estimate of the population employment for the frame quarter. Yet the ECEC mean wage and benefit estimates, and their microdata, are for the current quarter, which for this research is at least eleven quarters after the frame quarter. If the weights are unaltered, the weight distribution will match the population employment distribution for the frame quarter, not the current quarter. Sample attrition (non-response, and establishment death) also occurs within a group over time. If the weights of respondent units are unaltered, some of the total sample weight for the group will be lost, and since the loss is not uniform, this also alters the weight distribution. Note that the loss of sample weight due to death is acceptable, because this estimates the loss on the frame due to death. Yet the sample for a group is never supplemented, so any weight-gain caused by new births is not captured.

Two methods are used to realign the ECEC weight distribution: non-response adjustment and benchmarking. Non-response adjustment is done first and transfers weight from non-respondents to respondents. Hence, no weight is lost and the original weight distribution is preserved. Non-response adjustments are done independently by non-response adjustment cell, which are defined based on locality, industry sampling stratum, and establishment size class. Benchmarking is accomplished by multiplying the non-response-adjusted weight by a benchmark factor T/W, where T is the current population employment, and W is the sum of the non-response-adjusted weights. Benchmarking serves two purposes. First, it ensures that the weight distribution matches the population employment distribution of the current quarter, not the frame quarter. Second, for each rotation group, it captures the net impact of establishment turnover that occurs after the frame quarter. Benchmarking is done independently by benchmark cell. Currently, the benchmark cells are industry sampling strata.

2. Research Summary

ECEC estimates indicate that compensation costs are correlated with establishment size. Figure 1 shows mean wages and mean total benefit costs for four size classes, using the current benchmarking method. Costs tend to increase with establishment size.

Figure 1. Mean Hourly Wages and Total Benefit Costs, by Size Class, March 2012



This research studies the effect of splitting each industry benchmark cell further into establishment size classes. Seven benchmark methods are studied. The first uses the original, non-benchmarked weights. The second uses the current ECEC procedure, which benchmarks by industry. The last five methods benchmark by industry and size class. Each successive method splits the size classes into ever smaller cells. See Figure 3 in Section 5 for the cell definitions.

There are two research phases. In Phase 1, the existing NCS sample is used. Estimates of mean hourly cost and standard error are computed for each benchmark method, domain (population subset), and compensation component (wage, benefit, or aggregate). In Phase 2, the NCS samples are simulated. For each rotation group, 1000 new establishment samples are selected. For each sample, benchmark method, and domain, an estimate of mean monthly earnings is computed (the frame contains no data on benefits or hours-worked). This simulated distribution of 1000 estimates is then used to approximate the expected value, bias, standard error, and root mean squared error (MSE) for each estimator.

For both Phases, we expected that benchmarking by size class would cause a decrease in the mean cost, absolute bias, standard error, and root MSE. Also, the more size classes that we used, the larger the decrease. We expected costs to drop because small establishments

tend to have larger attrition rates, and hence larger benchmark factors, than large establishments. Small establishments also have lower mean costs. Hence the weight-share allocated to low-cost establishments should increase. If the costs drop, then their expected values will drop, too. We expected the absolute bias to fall, since we assumed finer cell definitions might improve the estimator, for the same reason stratified samples often do. We expected standard errors to drop, based on an investigation of formulas (see Sections 4 and 5). If the absolute bias and standard error drop, then the root MSE drops.

For Phase 1, the output met our expectations. See Section 6 for more details. On average, benchmarking by size class causes estimates of mean wage and total benefits to decrease. Standard errors also decrease, as expected. The more size classes we use, the larger the decline. For each domain and method, we subtract its estimate from the current method. About a quarter of these differences are statistically significant at the 90% confidence level.

For Phase 2, the output only partially met our expectations. See Section 7 for more details. On average, benchmarking by size class causes the expected value and absolute bias to increase. We expected them to drop. We do not yet know the reason for the increase. As expected, the standard error decreases, and the more size classes we use, the larger the drop. The root mean squared error increases at first, yet then decreases the more size classes we use, because the standard error is about twice the absolute bias, and drops more steadily.

3. The National Compensation Survey

The scope of the NCS includes the 50 states and the District of Columbia. Data are collected on the mean hourly employer-costs of wages and benefits. Respondents are also asked which benefit plans they offer, the percent of workers who participate, and the details of plan provisions. The NCS microdata support several products. The most important are:

- *Employment Cost Index*, ECI, which tracks changes in wages and benefit costs over time for a fixed market-basket of workers.
- *Employer Costs for Employee Compensation*, ECEC, which publishes mean wage and benefit costs for the current market-basket of workers.
- *Employee Benefits Data*, which includes a variety of statistics on benefits, including access and participation rates, and the detailed provisions of benefit plans.

Our research only focuses on the ECEC, specifically, the cell-definitions used to compute benchmark factors. Also, to reduce the workload, our research is restricted to private industry establishments. A more detailed description of the NCS sample design and ECEC weighting process is given in Chapter 8 of the BLS Handbook of Methods, and in McCarthy et al. (2011).

The NCS microdata used in this research comes from a three-stage sample design. In Stage 1, a sample of 152 localities is selected from a frame that spans the nation. In Stage 2, a sample of establishments is selected from each sampled locality. In Stage 3, a sample of jobs is selected from each respondent establishment. A job is a collection of workers in an establishment with the same attributes, such as occupation, work-level, union status, full-time vs. part-time status, and whether or not the worker receives incentive-pay. The work-level is defined by the BLS and is based on job duties and responsibilities, the work performed, and the skills, education, and training that are required. The NCS establishment sample is based on a rotating-group design. Typically, once a year, the oldest sample rotation group is dropped and replenished by a new rotation group.

The BLS Quarterly Census of Employment and Wages (QCEW) serves as the primary source of the establishment sampling frame for the NCS. 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). The NCS frame is further supplemented by railroad establishments, since the UI does not contain all railroad data. For each rotation group, its establishment sample is selected from the most recent NCS frame. For this research, the frame data are three to five quarters old by the time a group is sampled. Yet for the most recent rotation group (which was not used in this research), the frame data are only two quarters old.

Once a rotation group is sampled, it is initiated (its establishments are contacted for the first time). To spread out the workload, the establishments in a rotation group are split into four collection panels, and one panel is initiated each quarter. Once an establishment is initiated, its data are updated each subsequent quarter until the establishment drops out (stops participating, its jobs no longer exist, or its group rotates out). Once all four collection panels for a group are initiated, the group is updated one more quarter, and then establishment weights are computed for the group as a whole. These include adjustments for non-response. The new rotation group is first used in ECEC estimates the first quarter after weights are computed for the group.

Non-response adjustment is done by multiplying the original weight by a factor V/R, where V is the sum of the original weights of viable establishments, and R is the weight-sum of respondent establishments. A viable establishment is in scope, in business, and contains at least one viable job. Non-response adjustment is done independently by cell, which are of the form (locality × industry sampling strata × size class). The size classes are {1-49, 50-99, 100-499, 500+}, which are the same as those used in method B3 of our research. Non-response adjustment is done once for a group and the factors are fixed until the group rotates out.

ECEC estimates are weighted averages of the contributions of each active rotation group. Benchmark factors are recomputed each quarter. The numerator is the most recent employment target, the denominator uses all active rotation groups. Benchmark cells are industry sampling strata. Currently, geographic and size class breakouts are not used.

Figure 2 illustrates the private industry rotation group lifecycle. When this research began, the most recent frame available was 1st quarter 2012, so all estimates in this research are for 1st quarter 2012. The published private industry estimates for that quarter were based on rotation groups 105-109, so our research was restricted to those groups.

Figure 2. Private Industry Rotation Group Lifecycle

Group	2004		2005				2006				2007				2008				2009				2010				2011				2012	
	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2
105	F					S			Initiated					Used in Estimates																		
106					F					S			Initiated				Used in Estimates															
107									F					S			Initiated				Used in Estimates										
108														F				S			Initiated				Used in Estimates						
109																		F				S			Initiated				In Ests		
110																						F				S			Initiated			
111																											F			S		
112																															F	

F: The frame quarter. S: The sample was selected in this quarter, yet from the frame quarter F.

4. NCS Non-Benchmarked and Benchmarked Estimators

This research studies both non-benchmarked and benchmarked ECEC mean cost estimators. The non-benchmarked mean hourly cost for a domain (subset) *D* is given by

$$\hat{\bar{Y}}_D = \frac{\sum_{q \in D} \hat{W}_q \bar{Y}_q}{\sum_{q \in D} \hat{W}_q}$$

where

q is any usable job in the domain *D*.

$\hat{W}_q$ is the non-benchmarked weight of the job q

$\bar{Y}_q$ is the mean hourly cost of the job q

The benchmarked mean hourly cost for a domain *D* is given by

$$\hat{\bar{Y}}_{BD} = \frac{\sum_{q \in D} \hat{B}_q \hat{W}_q \bar{Y}_q}{\sum_{q \in D} \hat{B}_q \hat{W}_q}$$

where $\hat{W}_q$ and $\bar{Y}_q$ are as before, and $\hat{B}_q$ is the benchmark factor for the job q. Suppose job q is in benchmark cell *C*. Then the benchmark factor $\hat{B}_q = \hat{B}_C = T_C / \hat{W}_C$, where

T_C is the benchmark employment target for cell *C*.

$\hat{W}_C = \sum_{q \in C} \hat{W}_q$ is the non-benchmarked estimate of total employment for cell *C*.

The benchmark targets are derived from the most recent NCS frame. It takes several quarters for a frame to become available. For 1st quarter 2012 estimates the delay was 3 quarters. For more recent quarter estimates, the delay is only 2 quarters. To compensate for the time lag between the most recent frame quarter and the current quarter, we use data from the BLS Current Employment Survey (CES) to age the frame counts forward in time, using a process similar to benchmarking. For more details on the aging process, see McCarthy et al. (2011). For this research, however, we had the frame for the current quarter (1st quarter 2012), so there was no need to age the frame targets using the CES.

In some cases, the cells are so small that their estimates may be too unreliable. In other cases, the benchmark factors are extreme. Hence, some cells are collapsed into larger aggregates. In this case, the cells *C* in the formulas would be replaced with the final cell definitions after collapsing. For the published ECEC, the cells *C* (before collapsing) are the industry sampling strata used to select the establishment sample. For the 24 private industry cells, no collapsing (of industries) was required, so the final industry cells equal the original industry cells. For the 20 government cells, some collapsing was necessary. Yet this research only focused on private industry data.

The purpose of this research is to study the effects of subdividing the 24 private industry cells further into size classes. Some of these size class cells require cell-collapsing. Yet if a cell requires collapsing, it was first combined with other size classes within the same industry. If all size classes in an industry need collapsing, then the final cell is simply the industry.

To compare the non-benchmarked formula with the benchmarked formula, it is helpful to rewrite them as weighted averages of cell mean cost. The non-benchmarked mean cost for domain *D* can be written in this form:

$$\hat{\bar{Y}}_D = \frac{\sum_C (\sum_{q \in D \cap C} W_q Y_q)}{\sum_C (\sum_{q \in D \cap C} W_q)} = \frac{\sum_C \hat{Y}_{DC}}{\sum_C \hat{W}_{DC}} = \frac{\sum_C \hat{W}_C \left(\frac{\hat{W}_{DC}}{\hat{W}_C}\right)\left(\frac{\hat{Y}_{DC}}{\hat{W}_{DC}}\right)}{\sum_C \hat{W}_C \left(\frac{\hat{W}_{DC}}{\hat{W}_C}\right)} = \frac{\sum_C \hat{W}_C \hat{F}_{D|C} \hat{\bar{Y}}_{DC}}{\sum_C \hat{W}_C \hat{F}_{D|C}}$$

The benchmarked mean cost for domain *D* can be written in this form:

$$\hat{\bar{Y}}_{BD} = \frac{\sum_C B_C (\sum_{q \in D \cap C} W_q Y_q)}{\sum_C B_C (\sum_{q \in D \cap C} W_q)} = \frac{\sum_C B_C \hat{Y}_{DC}}{\sum_C B_C \hat{W}_{DC}} = \frac{\sum_C \left(\frac{T_C}{\hat{W}_C}\right) \hat{W}_{DC} \left(\frac{\hat{Y}_{DC}}{\hat{W}_{DC}}\right)}{\sum_C \left(\frac{T_C}{\hat{W}_C}\right) \hat{W}_{DC}} = \frac{\sum_C T_C \hat{F}_{D|C} \hat{\bar{Y}}_{DC}}{\sum_C T_C \hat{F}_{D|C}}$$

where:

$\hat{\bar{Y}}_{DC} = \hat{Y}_{DC} / \hat{W}_{DC}$ = non-benchmarked estimate of the mean hourly cost for workers in domain *D* and cell *C*

$\hat{F}_{D|C} = \hat{W}_{DC} / \hat{W}_C$ = non-benchmarked estimate of the fraction of workers in cell *C* that are in domain *D*

$\hat{Y}_{DC} = \sum_{q \in D \cap C} \hat{W}_q Y_q$ = non-benchmarked estimate of total hourly cost for workers in domain *D* and cell *C*

$\hat{W}_{DC} = \sum_{q \in D \cap C} \hat{W}_q$ = non-benchmarked estimate of total employment for workers in domain *D* and cell *C*

$\hat{W}_C = \sum_{q \in C} \hat{W}_q$ = non-benchmarked estimate of total employment for cell *C*

For many domains, the cell *C* is a subset of the domain *D*. If so, the fraction $\hat{F}_{D|C} = 1$.

The non-benchmarked and benchmarked mean cost formulas differ only in the cell-weights ($\hat{W}_C \hat{F}_{D|C}$ and $T_C \hat{F}_{D|C}$) used in the weighted average. Note that $\hat{W}_C$ can vary by sample, but T_C is fixed for all samples. This suggests that the variance of benchmarked estimators might tend to be lower than that of non-benchmarked estimators. But there are exceptions. For example, suppose a cell has a large mean cost variance, and its share of the total non-benchmarked employment for the domain is notably less than its share of the benchmarked employment for the domain. Then benchmarking may increase the relative impact of this high-variance cell, which may lead to a higher variance overall for the mean cost estimate for the domain.

5. Splitting Benchmark Cells into Size Classes

For Phases 1 and 2, seven different estimation methods were studied. See Figure 3. The label NB means non-benchmarked, B means benchmarked. The B0 method used the same 24 industry cells as the published ECEC. Methods B1-B5 used the same industry cells as B0, yet each industry cell was broken out further into size classes. The size classes for methods B1 and B3 are those we publish. Those for B3 are used to define non-response adjustment cells. Those for B5 correspond with size classes used in the QCEW program.

Figure 3. Estimators Studied, and Their Size Classes

Bench-marked?	Esti-mator	Size Classes Used by Estimator								
No	**NB**	All workers (no size class breakout)								
Yes	**B0**	All workers (no size class breakout)								
	B1	1-99 workers					100 or more workers			
	B2	1-99 workers					100-499 workers		500 or more	
	B3	1-49 workers				50-99	100-499 workers		500 or more	
	B4	1-49 workers				50-99	100-249	250-499	500 or more	
	B5	1-4	5-9	10-19	20-49	50-99	100-249	250-499	500-999	1000+

Benchmarking by size class should tend to lower the variance estimate further, and the smaller the size classes, the lower the variance, in general. For example, suppose we have two benchmarked estimators, A and B, where the cells of B are smaller than, and contained within, the cells of A. Then to compare the estimators, it helps to express them in terms of the smaller B-cells. For the following, let the symbol *A* refer a cell used in estimator A, and the symbol *B* refer to a cell used in estimator B.

Estimator A, with the larger cells, can be written as:

$$\frac{\sum_A \left(\frac{T_A}{W_A}\right)\left(\sum_{B\in A} \hat{W}_{DB} \bar{Y}_{DB}\right)}{\sum_A \left(\frac{T_A}{W_A}\right)\left(\sum_{B\in A} \hat{W}_{DB}\right)} = \frac{\sum_A \sum_{B\in A} T_A \left(\frac{\hat{W}_{DB}}{W_A}\right)\left(\frac{\bar{Y}_{DB}}{W_{DB}}\right)}{\sum_A \sum_{B\in A} T_A \left(\frac{\hat{W}_{DB}}{W_A}\right)} = \frac{\sum_A \sum_{B\in A} (T_A \hat{F}_{B|A}) \hat{F}_{D|B} \bar{Y}_{DB}}{\sum_A \sum_{B\in A} (T_A \hat{F}_{B|A}) \hat{F}_{D|B}}$$

Estimator B, with the smaller cells, can be written as:

$$\frac{\sum_A \sum_{B\in A} \left(\frac{T_B}{W_B}\right) \hat{W}_{DB} \bar{Y}_{DB}}{\sum_A \sum_{B\in A} \left(\frac{T_B}{W_B}\right) \hat{W}_{DB}} = \frac{\sum_A \sum_{B\in A} T_B \left(\frac{\hat{W}_{DB}}{W_B}\right)\left(\frac{\bar{Y}_{DB}}{W_{DB}}\right)}{\sum_A \sum_{B\in A} T_B \left(\frac{\hat{W}_{DB}}{W_B}\right)} = \frac{\sum_A \sum_{B\in A} (T_B) \hat{F}_{B|B} \bar{Y}_{DB}}{\sum_A \sum_{B\in A} (T_B) \hat{F}_{D|B}}$$

All variables are the same as in Section 4 (except the symbol *C* is replaced with *A* or *B*). The new term $\hat{F}_{B|A} = \hat{W}_B / \hat{W}_A$, and is equal to the non-benchmarked estimate of the fraction of workers in cell *A* that are in cell *B*.

Estimator A and Estimator B differ only in the terms in parentheses, $T_A \hat{F}_{B|A}$ and T_B. Note that $\hat{F}_{B|A}$ can vary by sample, but T_A and T_B are fixed for all samples. This suggests that the variance of Estimator B might tend to be lower than that of Estimator A. That is, we expect benchmarking further by size class will lower the variance, and the smaller the size classes, the lower the variance. Again, there are exceptions. Suppose we have a high-variance B-cell whose share of the denominator of estimator A is less than its share of the denominator of estimator B. Then benchmarking by the B-cells rather than A-cells may increase the relative impact of this high-variance cell, which may lead to a higher variance overall for the mean cost estimate for the domain.

6. Phase 1: Using One Real NCS Sample

For Phase 1, we use the same private industry microdata (mean costs $\bar{Y}_q$ and weights $\hat{W}_q$) that were used to produce the ECEC published values for 1st quarter 2012. A total of 142 estimation domains were studied, corresponding roughly to those that are currently published. For each domain, estimates are produced for 27 compensation components (total compensation, wages, total benefits, 6 benefit aggregates, and 18 individual benefits).

Figure 4 shows mean wages for the largest domain we studied: All Private Industry Workers. Figure 5 shows the mean cost of total benefits. The colored-bands are 90% confidence intervals. The radius of each interval is given by the green number above the interval. Standard errors are computed using balanced repeated replication. The benchmark factors are recomputed for each replicate.

Figure 4. Mean Wage, 90% Confidence Intervals, by Method, All Private Workers

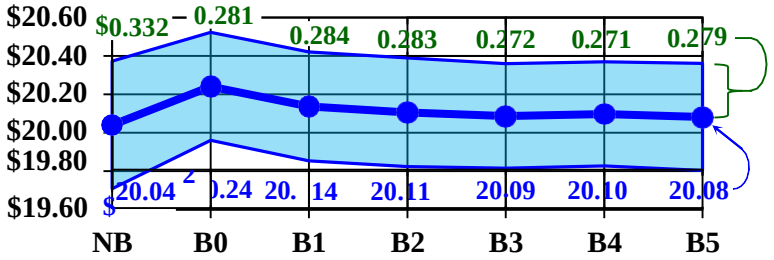
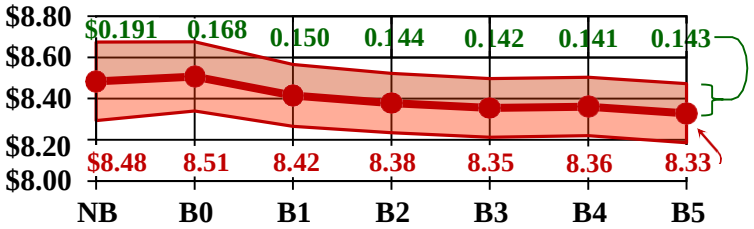


Figure 5. Mean Total Benefits, 90% Confidence Intervals, by Method, All Private



For All Private, the minimum mean costs occur for method B5. The wage for B5 is 16 cents less than the wage for B0, a drop of 0.8%. For total benefits, the decrease is 18 cents (2.1%). The standard error minimums occur for B4. The wage standard error for B4 is 0.6 cents less than B0, a drop of 3.6%. For total benefits, the drop is 1.7 cents (16%).

To gauge the effect of a method *M* for a single domain, we compute its percent difference from method B0, which is 100 × (*M* value – B0 value) / (B0 value). To see the global effect, we average these percent differences over all 142 domains. Figure 6 shows average percent differences for wages and total benefits. Benchmarking by size class tends to lower the estimates, and the more size classes, the lower the estimates tend to be, in general.

Figure 6. Average Percent Difference (from B0), for Mean Costs, by Method

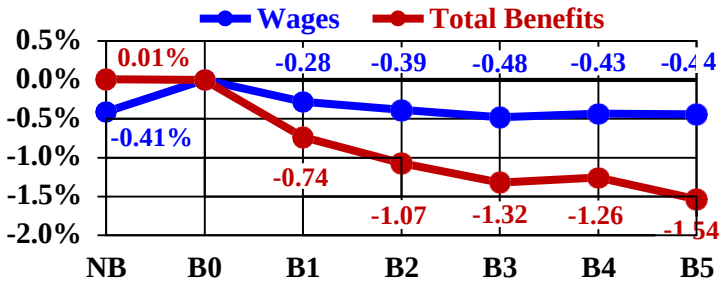
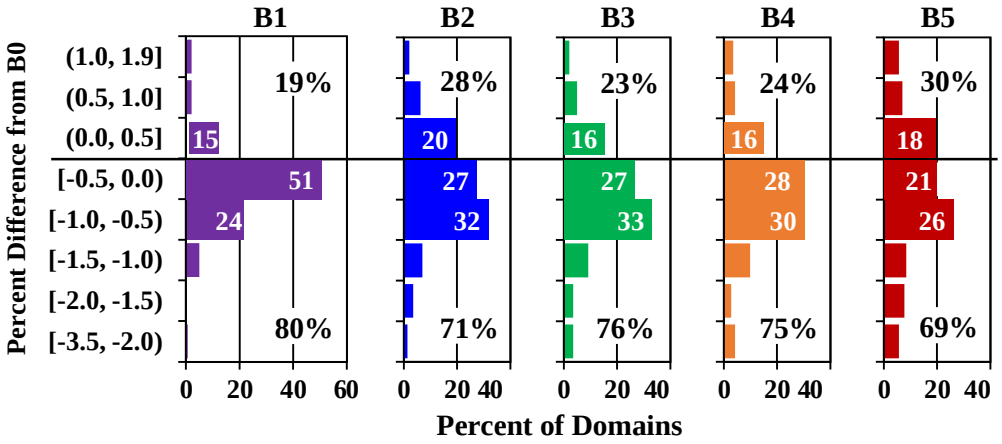


Figure 7 shows the distribution, across all 142 domains, of the percent differences for mean wages. The numbers on the top-right of each chart are the percent of domains with positive difference. The numbers on the bottom-right are the percent of domains with negative difference. One domain had zero difference for all methods, and so is excluded.

Figure 7. Distribution of Percent Differences (from B0), for Mean Wage



About 27% of the estimates for methods B1-B5 were significantly different from method B0, yet this varies by method and compensation component. The "All Components" row in Figure 8 shows the percent of estimates (across all components and domains) which are significantly different at the 90% confidence level. The other rows are restricted to one component (wages or total benefits). To estimate the variance of the difference, we let the replicate estimate of the difference equal the difference of the replicate estimates.

Figure 8. Percent of Estimates That Are Significantly Different From B0, by Method

Component	All Methods	B1	B2	B3	B4	B5
All Components	**27%**	26%	30%	29%	26%	23%
Wages	20%	22%	26%	21%	18%	15%
Total Benefits	39%	37%	43%	40%	38%	35%

For the standard errors, we first look at percent differences from B0. Figure 9 shows their averages for wages and total benefits. Figure 10 shows distributions of percent differences, for wage standard errors. Benchmarking by size class tends to lower standard errors.

Figure 9. Average Percent Difference (from B0), for Standard Errors, by Method

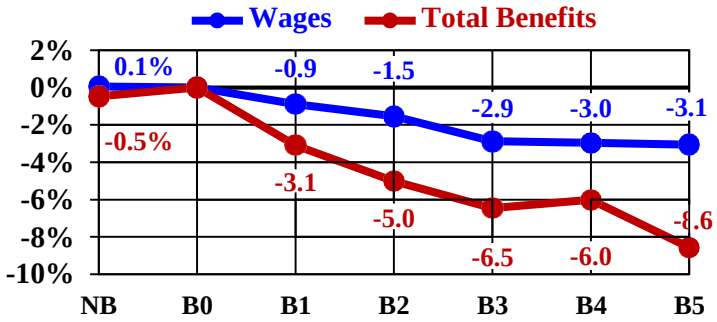
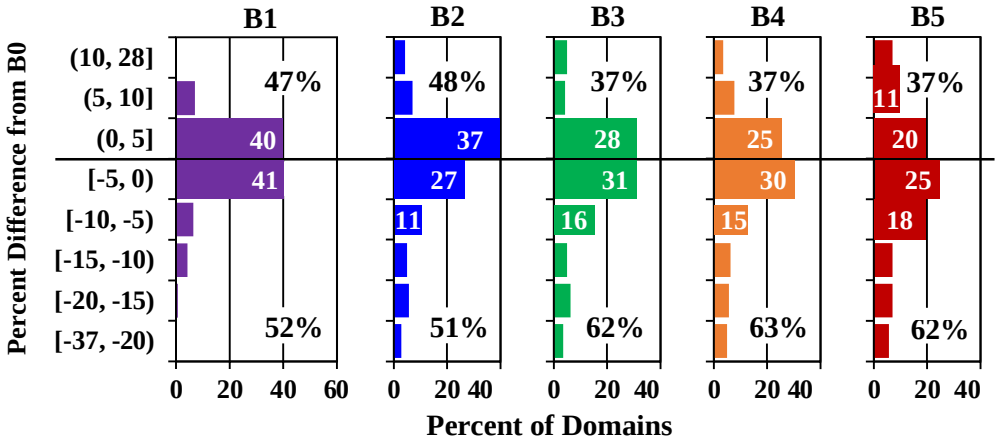
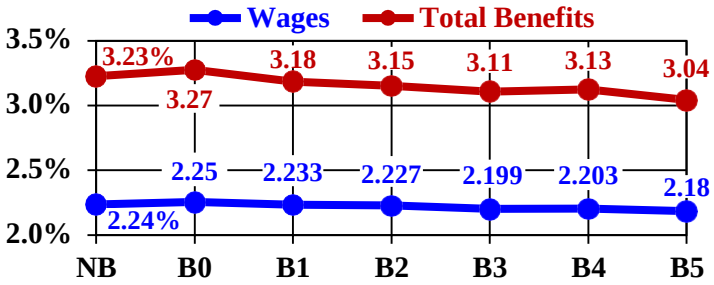


Figure 10. Distribution of Percent Differences (from B0), for Wage Standard Errors



Next, we focus on the percent relative standard error (%RSE), which is the standard error, as a percent of the estimate. Figure 11 shows average %RSEs. Benchmarking by industry tends to raise the %RSE slightly, yet benchmarking by size class tends to bring it down.

Figure 11. Average Percent Relative Standard Error, for Mean Costs, by Method



7. Phase 2: Using One Thousand Simulated NCS Samples

For Phase 2 we select 1000 simulated samples. We attempt to replicate, as close as possible, the sample design used to get Phase 1 microdata. Yet the availability of data, and the complexity of the sample design, places restrictions on the simulation. For the real NCS, a 3-stage design (sample localities, then establishments, and then jobs) was used. For the Phase 2 simulations, however, we can only take 1000 Stage 2 samples (of establishments).

For the Stage 1 sample, we reuse the single locality sample that is used in Phase 1. Although selecting 1000 new locality samples would not have been difficult, it would have been very difficult to replicate the sample-allocation process 1000 times (one allocation for each locality sample). The sample allocation process distributes the total establishment sample size to all the sampling strata (defined by rotation group, locality, and industry group). Hence, because we have only one allocation (the existing one), we only use one locality sample (the existing one).

For the Stage 2 samples, we assume all sampled units respond. We do, however, model attrition caused by establishments going out of business. If a sampled establishment fails to appear in any future frames, from the initiation quarter to the estimation quarter, then it

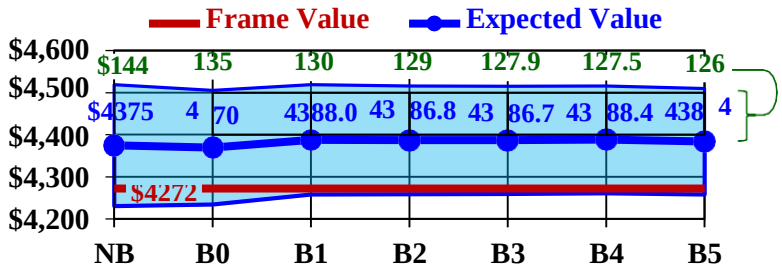
is dropped from the sample and not used in the estimates. See Figure 2 for the rotation group lifecycle. Note that in the real NCS, initiations are spread over 4 quarters and updates occurred every quarter. For Phase 2, however, we had only obtained frames for those quarters where samples are selected. In Figure 2, these correspond with those columns that contain a code "F" cell. Because of these gaps, initiations are condensed into one quarter (nearest F), and updates only occur about once a year (later F's). Hence the Phase 2 model may have slightly less attrition, caused by establishments going out of business, than in the real NCS. Yet since the real NCS has non-response, its attrition rates are higher overall.

Stage 3, the sampling of jobs, could not be implemented in Phase 2, because the establishment sampling frames do not contain data for individual jobs, and only have data for the establishment as a whole. For Phase 2, the microdata is restricted to mean monthly earnings, by establishment. This mean is derived from QCEW data, and is equal to the total quarterly earnings for the establishment, divided by the sum of its three monthly employment values for the quarter. Hence, the formulas in Section 3 must be modified. Jobs q are replaced with establishments k. Rather than using mean hourly job costs $\bar{Y}_q$ and job weights $\hat{W}_q$, for Phase 2 we use mean monthly establishment earnings $\bar{Y}_k$ and establishment weights $\hat{W}_k$. For Phase 2, we use the same benchmark targets we used for Phase 1.

For Phase 2, the estimation domains are similar to those that are published. Yet because Phase 2 is restricted to establishments, we do not have estimates for occupational groups or worker characteristics. For each domain, we have the frame value we are trying to estimate, and 1000 estimates of that value. Four statistics are approximated: the expected value, the bias, the standard error, and the root mean squared error. The expected value is the average of the 1000 estimates. The bias is the difference between the expected value and the frame value. The standard error is the square root of the variance, and the variance is the average squared deviation of the estimates from the expected value. The root mean squared error (MSE) is the square root of the MSE, and the MSE is the average squared deviation of the estimates from the frame value.

Figure 12 shows mean earnings for All Private Industry Workers. The red line is the frame value we are trying to estimate. The blue line shows the expected values. The blue band shows the 90% confidence intervals. The green number on top is the radius of the interval.

Figure 12. Expected Values, Frame Value, 90% Confidence Intervals, All Private

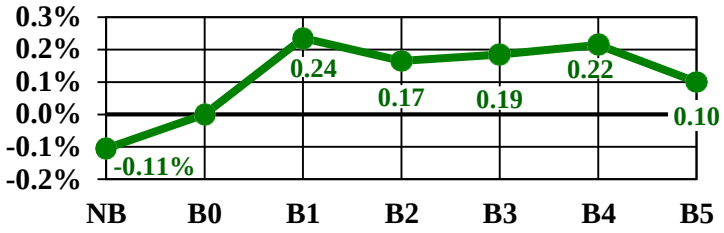


For All Private, the expected values and absolute biases for B1-B5 are all larger than that of B0, and the maximum is at B4. The expected value for B4 is $19 more than the expected value for B0, an increase of 0.4%. The absolute bias is also $19 more, yet since the bias is smaller than the expected value, the increase was 19%. These increases are unexpected,

and we do not yet know the reasons why they occur. For methods B1-B5, the standard errors are all less than that of B0, and the minimum is for method B5. The standard error for method B5 is $5.70 less than that of method B0, a drop of 6.9%. These decreases are expected. The root MSEs for methods B1-B5 are all larger than that of B0, and the maximum occurs for method B1, whose root MSE is $12.90 more than B0, a 10% increase. Yet method B5 has the lowest increase, at only $8, or 6.3%.

To compare expected values between methods, we look at percent differences from B0. Then we average across all 63 domains. See Figure 13. Benchmarking by industry tends to raise the expected value, yet benchmarking by size has little effect until method B5.

Figure 13. Average Percent Difference (from B0), for Expected Values, by Method



For absolute bias, some percent differences are extreme. So instead we compute the log-relative of method *M* with respect to B0, equal to *ln*(*M* value / B0 value). Figure 14 shows average log-relatives for absolute bias. The maximum is 0.26 and occurs for method B1 (its average percent difference is 30%). Figure 15 shows the distributions of log-relatives.

Figure 14. Average Log-Relative (with respect to B0), for Absolute Bias, by Method

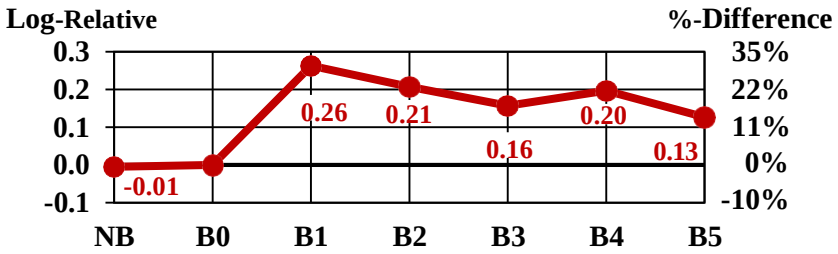
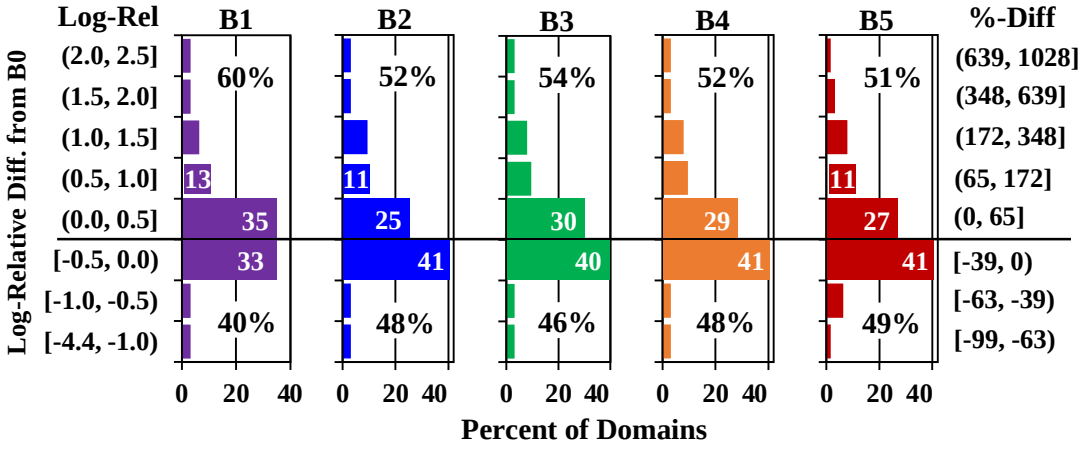


Figure 15. Distribution of Log-Relatives (with respect to B0), for Absolute Biases



For the standard errors and root MSEs, we first focus on their average percent differences from method B0. See Figure 16. Benchmarking by size class tends to lower the standard error. Since bias increases, however, benchmarking by size class raises the root MSE for method B1. As more sizes classes are used, however, the bias levels off, and the standard error continues to drop. Since the standard error is about twice the bias, the root MSE starts to drop for B2-B5. Figure 17 shows the distribution of percent differences, for the standard error. Figure 18 shows the distributions for the root mean squared error.

Figure 16. Average Percent Difference (from B0), for Standard Error and Root MSE

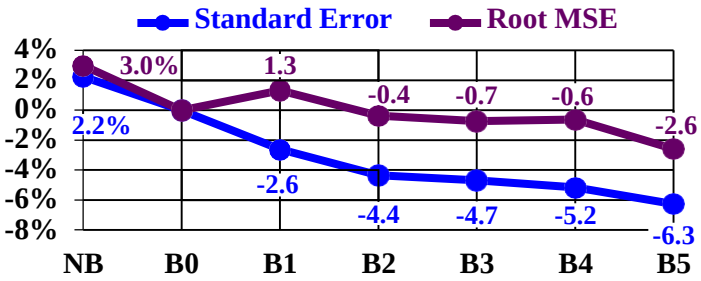


Figure 17. Distribution of Percent Differences (from B0), for Standard Errors

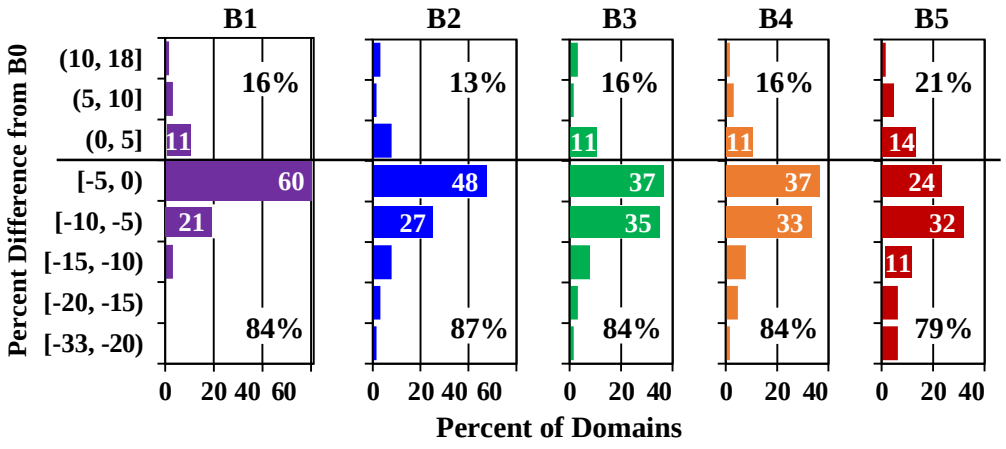
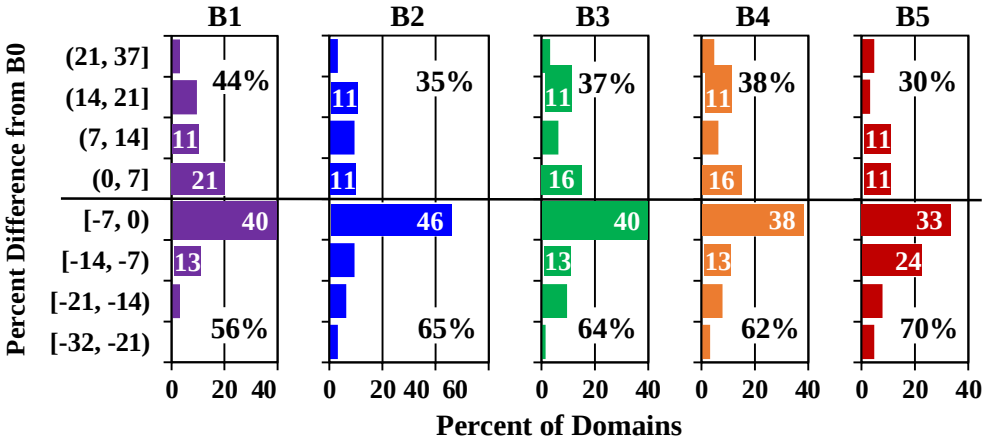


Figure 18. Distribution of Percent Differences (from B0), Root Mean Squared Error



Next, we focus on three alternate reliability measures. First, is the percent relative absolute bias, which is the absolute bias, expressed as a percent of the frame value. Second, is the percent relative standard error, which is the standard error, expressed as a percent of the expected value. Third, is the percent relative root MSE, which is the root MSE, expressed as a percent of the frame value. Figure 19 shows their averages over all domains.

Figure 19. Average Percent Relative Absolute Bias, Standard Error, and Root MSE

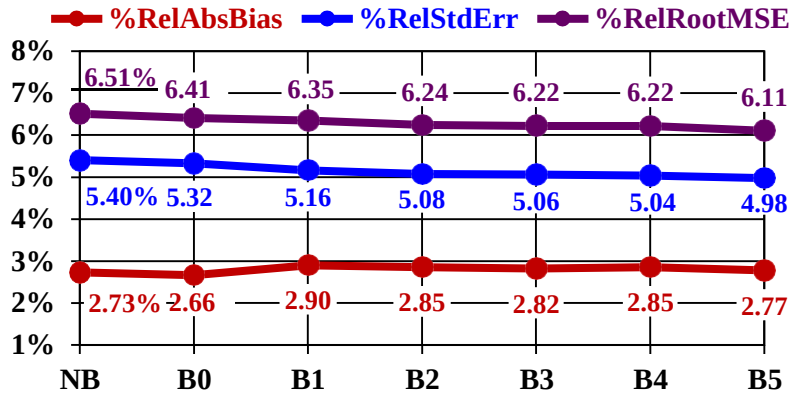
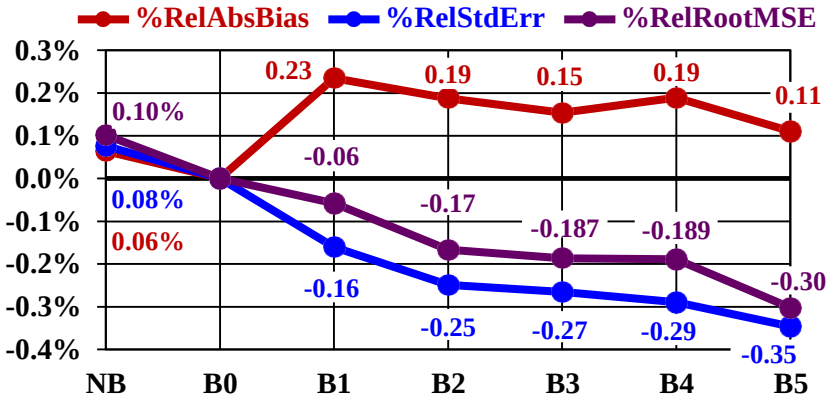


Figure 20 shows their average differences from method B0. We look at differences, rather than percent differences, since the values are already represented as percentages.

Figure 20. Average Differences (from B0), for Percent Relative Statistics



8. Conclusion

Benchmarking by size class tends to lower the standard errors for both Phases, and lower the root mean squared error for Phase 2. These results were expected. The more size classes we used, the lower the standard error and root MSE, in general. Method B5, which used the most size classes, produced the lowest standard errors and root MSEs, on average.

We expected estimates to tend to decline when we benchmarked by size class. In Phase 1, estimates do decrease, on average, yet for Phase 2, the expected values of the estimates increased. We do not yet know the reason for the increase. For Phase 2, we expected the absolute bias to also drop, yet instead it rose. Again, we do not yet know the cause. But the

more sizes classes we used, the absolute bias came back down (yet still was more than method B0, on average).

There was a lot of variation among domains. For example, for many domains, standard errors rose and/or absolute bias dropped. So averages do not tell the whole story. Most of the analyses above focus on percent differences from B0, or differences for B0. This was done to allow comparison across method and domains. Yet the underlying changes were rather small, in an absolute sense. For example, for the largest domain, the Phase 1 standard error for wage only declined 16 cents, and for total benefits, 18 cents. Yet 27% of all new wage estimates (using methods B1-B5) were significantly different from method B0, at the 90% confidence level.

There are still unanswered questions and avenues for further research:

1. So far, our primary focus has been on the *average* effect of each method on domain-statistics (such as mean costs, bias, and variance). Less time has been spent, however, on studying the effects on individual domains or the impacts of each benchmark cell. We would like to know what is driving these results, and whether or not these effects and impacts are concentrated in certain types of domains and cells.

2. For Phase 2 we expected the estimates and absolute bias to decrease (on average) when we benchmarked by size class. Instead, they increased. We would like to study which benchmark cells had the most impact, to try and find the root cause for the increase.

3. Our research used microdata for an older quarter, whose contributing sample groups all came from an older, 3-stage sample design. We would like to extend our research to more recent quarters, including those quarters that contain some sample groups from the current, 2-stage sample design.

References

McCarthy, C. L., Ferguson, G. R., and Ponikowski, C. H.(2011), "The Weighting Process Used in the Employer Costs for Employee Compensation Series for the National Compensation Survey", *2011 Proceedings of the Section on Survey Research Methods*, [CD-ROM], Alexandria, VA: American Statistical Association.

U.S. Bureau of Labor Statistics (2013) *BLS Handbook of Methods*,
Chapter 5: Employment and Wages Covered by Unemployment Insurance,
http://www.bls.gov/opub/hom/homch5_a.htm
Chapter 8: National Compensation Measures,
http://www.bls.gov/opub/hom/homch8.htm

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