IMPUTATION OF BENEFIT RELATED DATA FOR THE NATIONAL COMPENSATION SURVEY

James A. Buszuwski, Daniel J. Elmore, Lawrence R. Ernst, Michael K. Lettau,
Lowell G. Mason, Steven P. Paben, Chester H. Ponikowski
buszuwski.james@bls.gov, elmore.daniel@bls.gov, ernst.lawrence@bls.gov, lettau.michael@bls.gov,
mason.lowell@bls.gov, paben.steven@bls.gov, ponikowski.chester@bls.gov
Bureau of Labor Statistics, 2 Massachusetts Ave. N.E., Room 3160, Washington, DC 20212-0001

KEY WORDS: missing data, item nonresponse, hot-deck, regression

1. Introduction

The National Compensation Survey (NCS) is an establishment survey of wages and employer-provided benefits conducted by the Bureau of Labor Statistics. It is the combination of three previously separate surveys: the Employer Cost Index (ECI), which measures quarterly and annual changes in wages and benefit costs and wage and benefit cost levels; the Employee Benefit Survey (EBS), which measures benefit incidence and provisions; and the locality wage survey, which measures wages for a sample of localities and for the nation as a whole. In addition to the continued publication of these surveys, new products linking benefit costs and provisions will be published as part of the NCS.

The integrated NCS sample is selected using a three-stage stratified design with probability proportionate to employment sampling at each stage. The first stage of sample selection is a probability sample of areas; the second stage is a probability sample of establishments within sampled areas; and the third stage is a probability sample of occupations within sampled areas and establishments. More details on the integrated sample is provided in Ernst et al. (2002).

The integration of these different surveys requires reconciling the different methodologies used in imputing for missing data items when they were separate. This paper describes the methods that will be used for imputing benefit incidence, provisions, time, and costs (for wage imputation, see Barsky et al. (2000). Section two describes the nearest neighbor methods used for imputing incidence and plan provisions, while sections three and four detail the regression models for imputing time and cost, respectively.

2. Benefit Incidence and Plan Provisions Imputation

In NCS, some respondents only provide wage information even though they were selected to provide both wage and benefit information. For the ECI, missing benefit costs for quotes with no collected benefit information were imputed so that each quote had a measure of total compensation. However, since the EBS was considered to be a separate survey pertaining solely to benefits, it had a separate weighting adjustment in order to account for these benefit nonrespondents. In order to link the incidence of benefits and the plan provisions to benefit cost data, we needed to extend the imputation of incidence of benefits and plan provisions to include benefit nonrespondents as a substitute for the benefits nonresponse adjustment.

In order to impute benefits data that was consistent across a plan, a study by Elmore, Loewenstein, and Buszuwski (2001) recognized that the order of the imputations was going to be important. First, whether or not at least one plan exists (i.e. incidence) in each benefit area for benefit nonrespondents needed to be determined. If no plan exists, then no further imputation is necessary. If a plan does exist, then the number of plans available to each sampled occupation needed to be imputed. This study proceeded with imputing the key provisions of each plan, since plan provisions are likely to influence other items, such as participation rates and employer costs. Next, the participation rate for each plan is imputed, since it was desired to take participation into consideration for the cost imputation. Finally, the remaining types of benefit data were imputed, which include: detailed provisions, employee and employer premiums for health plans, and time and employer benefit costs. This completes the imputation at initiation. During update collection, the imputed value for any missing data except cost data is generally the value for that item for the prior collection period, even if that value had been imputed. The imputation of cost data at update is described in Section 4.

Using the sequential approach, all of the benefits data for benefit nonrespondents ends up being imputed, as does any missing data for benefits respondent quotes. A benefits respondent quote is a quote for which the incidence and number of plans for each benefit area for the quote is known. All other quotes are benefits nonrespondent quotes. This led to concerns about not preserving relationships between benefit areas, and between types of data within a benefit, such as relationships between participation rates, provisions, and premiums of a health plan. The simplest solution to preserve these relationships for the benefit nonrespondents was to impute in the initial imputation as much data as possible from the same donor using a hot-deck method, with all benefits nonrespondent quotes being recipients in this imputation and all benefits respondent quotes eligible to be donors.

We chose to use a random within-cell hot-deck method for imputing missing benefit data. A prior study that compared imputation methods using EBS data showed that a random within-cell hot-deck method performs equally well as other methods (Montaquila and Ponikowski, 1993) In a random within-cell hot-deck method, imputation classes (or "cells") are formed, based on auxiliary data that is known for all units. Within each cell, the "unusable" unit (or "recipient") takes on the characteristic or characteristics of interest from a "usable" unit (or "donor") that is selected at random within the same cell. In our case, the missing benefit data for a sampled occupation (or quote) would take on all of a donor's usable data for each benefit area. That is,

$$y_{ij}^{*} = y_{ik},$$

where $y_{ij}^{*} =$ imputed usable data of each benefit area for quote *j* in cell *i*, and $y_{ik} =$ actual usable data of each benefit area for quote *k* in cell *i* such that quote *k* is chosen at random from among all usable quotes in cell *i*.

The cells are formed based on establishment characteristics and occupational characteristics. The cell is defined by the following auxiliary variables:

1. ownership (private or public sector)
2. size class
3. major industry division
4. major occupational group
5. two-digit NAICS code
6. union/non-union status
7. full-time/part-time status
8. Census region

If a suitable donor cannot be found for a particular cross-product of all cell variables, then the cell is collapsed one level and the process is repeated until a suitable donor is found. For example, if a suitable donor cannot be found for a recipient using cell variables 1-8, then census region is dropped and the cell would be collapsed to include only cell variables 1-7.

Additionally, we restrict the number of times a donor may be used to three. Previously when using hot-deck methods in EBS, we restricted the number of times a donor may be used to one. However, our research showed that by increasing the number of times to three, we could greatly reduce the amount of cell collapsing performed in the imputations. This should help to further reduce the bias albeit at the sacrifice of perhaps some variance increase.

Since we require that the incidence and number of plans for each benefit area for a donor be known, at a minimum this information will be assigned for the benefit nonrespondents after this initial imputation. (The number of plans available to a particular occupation may be greater than one for only seven benefit areas including: nonproduction bonus, life insurance, health insurance, short-term disability, long-term disability, defined benefit, and defined contribution. We assume the number of plans for the other benefit areas is equal to one.) However, any additional response data from the donor will also be known for a benefits nonrespondent, but any types of benefit data that are missing from the donor will remain missing from the recipient after this imputation. For example, a donor in this incidence and number of plans imputation may have missing participation rates for its health insurance plans. In this case, the imputed missing items of the benefit nonrespondents are maintained as missing after this imputation and will be handled in the subsequent imputations together with the corresponding missing items from the benefits respondents.

Next, we impute the key provisions. Key provisions are a select number of plan provisions for each benefit area that include plan identifying characteristics and are collected directly from the respondent. The key provisions in NCS are quite different from EBS, and the number of key provisions has increased. In EBS, imputation of the key provisions was not necessary. The difficulty with imputing for the key provisions is that the number and type of plan provisions are different for each benefit area. For example, there is only one continuous variable to impute for holiday leave, while there are 10 binomial variables for health insurance. Additionally, for some benefit areas there are many different combinations of missing and nonmissing key provisions. In order to preserve the relationships between provisions within each benefit area and preserve as much nonmissing data as possible, we decided to divide the provisions into sections and impute depending on the missing/nonmissing status of each section using a hot-deck method. The same donor would be used to replace all missing provisions sections on a recipient. Therefore, we have:

$$y_{ip}^{*} = y_{iq},$$

where $y_{ip}^{*} =$ imputed key provisions for plan *p* in cell *i*,

and $y_{iq} =$ actual key provisions for plan *q* in cell *i*, and plan *q* is chosen from among all plans with usable key provisions such that $|e_{ip} - e_{iq}|$ is minimized, where $e_{ip}, e_{iq} =$ establishment employment for establishment corresponding to plans p and *q* respectively in cell *i*.

Furthermore, some of the most important key provisions (generally referred to as "primary" provisions) are used in the cell definition for each benefit area such that at least some basic plan information is taken into consideration where possible. If a primary provision is missing, then that part of the cell is collapsed and a donor is determined strictly by establishment, occupational information, and any other primary provisions that may not be missing.

For example, the data collector should answer the following questions for health insurance plans:

1. Does the plan provide any of the following:
   1a. medical coverage?
   1b. dental coverage?
   1c. vision coverage?
   1d. prescription drug coverage?
2. Is the plan prepaid?
   2a. If yes, are there any restrictions on the choice of plan providers?
   2b. If no, can the enrollee go outside the network of plan providers for coverage at higher cost?
3. Does the employer pay any portion of claims for benefits?
   3a. If yes, is there a third party administrator?
   3b. If yes, is there a stop loss limit?

The answers for all of these questions are either yes, no, or unknown. However, for questions 1a-1d, either all are known or all are unknown. We divided the imputation for the three main health questions into three sections and impute based on the missing/nonmissing status of each section. For example, if question 3a is known and 3b is unknown, then the response for both 3a and 3b are imputed. Additionally, questions 1 (that is, 1a-1d), 2, and 3 are the primary provisions for health insurance. If any of these primary provisions are known, then this plan information is taken into consideration in addition to the establishment and occupational information in determining the donor.

The remaining benefit item that needs to be imputed prior to cost data is participation. Participation rates are collected for plans in six benefit areas: life insurance, health insurance, short-term disability, long-term disability, defined benefit, and defined contribution. Participation was previously collected in EBS and we used a nearest-neighbor within-cell hot-deck method to impute missing participation. After evaluating what was done in EBS, we believed there was little need for any major methodological changes for participation in NCS. However, we did have to make some minor changes to the cell definition necessitated by the changes in the collection and imputation of the key provisions. Therefore, the cells are now defined by establishment, occupational, and actual or imputed plan characteristics.

For participation, we simply have the following:

$$y_{ip}^* = y_{iq},$$

where $y_{ip}^* =$ imputed participation for plan *p* in cell *i*,

and $y_{iq} =$ actual participation for plan *q* in cell *i*, and plan *q* is chosen from among all plans with usable participation such that $| e_{ip} - e_{iq} |$ is minimized, where $e_{ip}, e_{iq} =$ establishment employment for establishment corresponding to plans p and *q* respectively in cell *i*.

Furthermore, a series of edit constraints are imposed upon participation data. For example, in health insurance, for a given occupation within a given establishment, the total participation rate cannot exceed 100% of the occupation employment separately by medical, dental, or vision coverage. As a result, imputed participation rates are sometimes modified in order to satisfy the edit constraint.

Other than time and cost data, the remaining benefit items to be imputed are the detailed provisions and the employee and employer health premiums. The order of these imputations is not important as long as they are after the imputation of key provisions. However, since they also use a nearest neighbor within-cell hot-deck method, we will address them here prior to addressing the time and cost imputations.

Detailed provisions are provisions collected from plan brochures or summary descriptions that describe all of the particulars of a plan. Detailed provisions are collected only for health insurance, defined benefit, and defined contribution plans. For detailed provisions, we do not impute by individual provisions or by sections of provisions as we did for the key provisions. Instead, we consider the situation as a binomial outcome, either the plan brochure is missing or nonmissing. If the plan brochure is missing, we impute the plan identifier of a plan with a nonmissing brochure and all of its detailed provisions are used for the missing brochure. In other words,

$$y_{ip}^* = y_{iq},$$

where $y_{ip}^* =$ imputed detailed provisions for plan *p* in cell *i*,

and $y_{iq} =$ actual detailed provisions for plan *q* in cell *i*, and plan *q* is chosen from among all plans with usable detailed provisions such that $| e_{ip} - e_{iq} |$ is minimized, where $e_{ip}, e_{iq} =$ establishment employment for establishment corresponding to plans p and *q* respectively in cell *i*.

The cells for the detailed provisions are formed based on establishment and occupational characteristics, and on a subset of the key provisions for each benefit area. They are defined as follows:

1. ownership (private or public sector)
2. benefit area
3. subset of key provisions for each benefit area

4. size class
5. major industry division
6. major occupational group
7. two-digit NAICS code
8. union/non-union status
9. full-time/part-time status
10. Census region

If a suitable donor cannot be found for a particular cross-product of all cell variables, then the cell is collapsed one level at a time until the cell is defined by only (1) ownership, (2) benefit area, and the (3) subset of key provisions. It is not permissible to collapse over cell variables 1-3 in the imputation of the detailed provisions. If the cell has been collapsed over cell variables 4-10 and a suitable donor has still not been found, then the restriction that a donor may be used only three times is relaxed to any number of times. The key provisions used for the imputation of the health insurance detailed provisions are coverage type (medical, dental, vision, prescription drug), indemnity or prepaid plan, and restrictions on the choice of providers or the ability of the participant to go outside the network. The key provision used for the imputation of defined benefit plans is employee contributory status to the plan. The key provision for the imputation of defined contribution plans is plan type (savings & thrift, money purchase, deferred profit sharing plan, etc.).

New to NCS is the estimation of employer health premiums. Previously in EBS, we only estimated employee health premiums. In ECI, employer premiums may have been collected as a part of determining the employer cost for health, which depends on rate and usage, so the cost is per employee. Here we are interested in the cost per participant. One of the characteristics we are interested in estimating is the relationship between employee and employer health premiums. Therefore, we believed it was important to take this relationship into consideration during imputation. If both the employee and the employer premium are missing, we impute the premiums simultaneously from the same donor. If only one premium is missing, then we use the non-missing premium as the nearest neighbor variable within the cell. In other words,

if both premiums are missing, then:

$y^*_{ipe} = y_{iqe}$, and $y^*_{ipr} = y_{iqr}$,

where $y^*_{ipe} =$ imputed employee premium for plan p in cell i,

$y^*_{ipr} =$ imputed employer premium for plan p in cell i,

$y_{iqe} =$ actual employee premium for plan q in cell i,

$y_{iqr} =$ actual employer premium for plan q in cell i,

and y_{iqe} and y_{iqr} are chosen from among all plans with usable premiums such that $|e_{ip} - e_{iq}|$ is minimized, where $e_{ip}, e_{iq} =$ establishment employment for establishment corresponding to plans p and q respectively in cell i.

if only the employee premium is missing, then:

$y^*_{ipe} = y_{iqe}$,

where y_{iqe} is chosen from among all plans with usable premiums such that $|y_{ipr} - y_{iqr}|$ is minimized,

if only the employer premium is missing, then:

$y^*_{ipr} = y_{iqr}$,

where y_{iqr} is chosen from among all plans with usable premiums such that $|y_{ipe} - y_{iqe}|$ is minimized.

The above imputation is done separately for single and family coverages. Therefore, we additionally require a donor to have nonmissing single employee and employer premiums, and nonmissing family employee and employer premiums. This guarantees that all imputed premiums for a recipient will be imputed based on premiums from the same donor plan.

3. Paid and Unpaid Time Imputation

The next part of benefits imputation is time, an umbrella term covering the average hours of overtime worked by employees within a selected occupation, as well as paid and unpaid weeks of vacation, and paid and unpaid days for holiday leave, sick leave, and other leave. These benefit areas are known as the time-based benefit areas.

For overtime, sick leave, and other leave, time used is collected. That is, if an occupation has one employee who is allowed 104 hours of sick leave per year, but that employee only takes 24, then it is the latter figure that is collected. For vacation and holiday time, it is assumed that everyone will (eventually) use up all of their available time, so time available is collected. That is, if the lone incumbent is allowed two weeks of vacation but only takes one, then it is the former figure that is collected.

For occupations that have been surveyed in a prior quarter, missing values of the time variables are filled in using values from the previous quarter, even if they themselves were imputed. The procedures in the remainder of this section apply only when no prior data is available.

The entire process consists of three steps: first, determine whether or not the occupation has access to the time-based benefit area in question; second, assuming it does have access, determine whether the plan offers paid or unpaid time; third, impute the amount of time (paid or unpaid) available or used per person in the occupation.

The first step is handled by the nearest-neighbor incidence model (see previous section). If the occupation in question is determined not to have access to a given time-based benefit area, then that observation is eliminated from the data set used throughout the remainder of the procedure for that benefit area. Otherwise, it's included.

Next, a binomial logistic regression is used to determine whether the plan being imputed offers paid or unpaid time. It is possible, though rare, for a plan to have both paid and unpaid time. In addition, some plans are collected as not having either paid or unpaid time. The latter happens most often in sick leave and other leave, where time used is collected rather than time available. That's because an occupation where, for example, none of the employees have taken any sick leave will show up as having zero days of both paid and unpaid sick leave. While the first iteration of the new system will never impute plans as having either both paid and unpaid time or neither, methods for allowing future versions to do so are under consideration. For now, such plans are not used as donors in this step.

The remaining donor plans from the benefit area in question are used in a binomial logistic regression where the dependent variable is a dummy coded 1 if the plan is paid and 0 if it is unpaid. The explanatory variables in the regressions are characteristics of the quote or establishment: the log average hourly wage, dummy variables for industry, occupation, establishment size, census division, full-time/part-time, union/nonunion, time/incentive paid, and ownership (i.e., private industry, local government, or state government), as well as the average annual work hours scheduled per person in the occupation. The model is used to generate the estimated probability that a given occupation with missing time data has a paid plan. A random number uniformly distributed between zero and one is then drawn. If the number is less than the estimated probability, the plan is imputed to have paid time. Otherwise, it's imputed to have unpaid time. There are at least two circumstance in which the model would fail to generate an estimated probability for one or more observations. First, the likelihood function could fail to converge, in which case no probabilities would be generated. However, this seems unlikely for such a simple regression. Second, it could be that all observations from a given industry, occupation, etc. could have missing values of the dependent variable. This would result in just those observations not having estimated probabilities. In any case, if the model fails to generate an estimated probability, a backup probability is generated from the total sample of good observations. That is, if 70% of all good observations from the benefit area in question are paid plans, then the predicted probability is set equal to .7, and the random number is compared against that. There is one such regression of each of the time-based benefit areas, except for overtime.

Finally, the number of hours, paid or unpaid, must be imputed. Five regressions, identical except for their dependent variable, are used to impute paid time, and four are used to impute unpaid time. (Overtime is always paid.) The dependent variable is the amount of time off provided by a plan, measured in hours for overtime, weeks for vacation, and days for holidays, sick leave, and other leave. The independent variables include those used in the logistic regressions, plus variables representing the amount of time off of the same type (i.e., paid or unpaid) as is being imputed provided by the plans in the other time-based benefit areas. Thus, in the regression for paid vacation weeks, the average number of overtime hours, paid holidays, paid sick leave days, and paid other leave days appear as independent variables. Similarly, in the regression for unpaid vacation weeks, the average number of unpaid holidays, unpaid sick leave days, and unpaid other leave days appear. If any of these variables are themselves missing, their value is set equal to zero. For each of these variables, a dummy variable is also included, coded 1 if the corresponding time variable has a missing value and zero if a value is present. These dummies can be thought of as measuring the average impact of paid or unpaid time from one benefit area being missing on how much time is offered in another benefit area. Thus, if the respondent being unwilling to furnish data for one time variable is correlated with high or low values of other time variables, this information will be incorporated in the model. This approach also prevents observations where, for example, paid weeks of vacation is available but paid days of sick leave are missing from being excluded from the regression for paid vacation.

Each observation in the regressions is weighted by its occupational sample weight. The fitted values generated by the models are used as the imputed values for occupations where the collected value was missing.

4. Benefit Cost Imputation

The imputation of benefit costs is different at initiation and update. We first explain the approach at initiation. NCS collects cost information for a benefit plan as the cost per employee in the quote. For example, suppose a quote has ten employees, eight of whom participate in a particular health insurance plan. If the cost among the eight workers is $10.00 per hour worked, the cost per employee in the quote will be $8.00. Thus, all else equal, a plan in which 80 percent of the workers in the quote participate will have a four times greater cost than a plan in which 20 percent participate.

The imputation procedure therefore first divides the cost per employee by the plan's participation rate. This converts the dependent variable in the regression to a cost per participant, which is a more homogeneous measure. The regressions for all the benefit areas have the following form.

$$C_{ij} = \alpha + \omega\ln(W_j) + \left(\sum_{k=2}^{9}\beta_k MID_{kj}\right) + \left(\sum_{k=2}^{9}\gamma_k MOG_{kj}\right) + \left(\sum_{k=2}^{4}\delta_k S_{kj}\right) + \left(\sum_{k=2}^{9}\lambda_k A_{kj}\right) + \chi FTPT_j + \phi U_j + \eta TI_j + \left(\sum_{k=1}^{K}\kappa_k KP_{kij}\right) + \varepsilon_{ij}$$

The subscript *i* is the index for plans and the subscript *j* is the index for quotes, so C_{ij} is the cost per participant for plan *i* in quote *j*. The set of key provisions, denoted individually by KP_{kij}, depend on the particular benefit area. As an example, for defined contribution pension plans, the key provisions are dummy variables for each of four types of defined contribution plans. The remaining explanatory variables in the regressions are characteristics of the quote or establishment: the log average hourly earnings, dummy variables for industry, occupation, establishment size, area, full-time/part-time, union/nonunion, and time/incentive paid. The set of donors is trimmed for the particularly high costs, as the top one percent of donors is excluded from the regressions.

The coefficients in the regression are estimated using the sample weight for the quote multiplied by the participation rate for the plan. The regression gives a fitted value, which is then multiplied by the actual or imputed participation rate for the plan to convert it back to the cost per employee. Predicting the cost per participant, then converting it to a cost per employee, ensures that the cost per employee for the plan is consistent with the plan's participation rate. Moreover, using the key provisions as explanatory variables ensures that the imputed costs are consistent with the characteristics of the benefit plan, regardless of whether they are collected or imputed.

One drawback to the imputation procedure is that donors need good participation data along with good cost data to allow the first-step conversion to the cost per participant. An alternative is to omit the first step and keep the dependent variable as a cost per employee. In this alternative, the number of plans for the quote is added as an explanatory variable. For example, suppose a quote has three health plans. Participation in the plans is 30, 20, and 50 percent, respectively, and the cost per employee is \$0.30, \$0.16, and \$0.55. The proposed imputation procedure uses \$1.00, \$0.80, and \$1.10 as the dependent variable because they equal the cost divided by the corresponding participation rate. The alternative uses \$0.30, \$0.16, and \$0.55 as the dependent variable, and adds three as an explanatory variable because, presumably, a larger the number of plans gives a lower the cost per employee for each plan.

The advantage to the alternative is that it does not require the participation data, so the pool of donors is potentially larger. Indeed, the alternative does give a roughly ten percent larger set of donors. However, the proposed procedure still does better in terms of the average absolute deviation of the out-of-sample predictions. The proposed procedure also does better with the out-of-sample prediction of the average level of the cost per employee among the benefit plans.

Establishments are scheduled to remain in the NCS for four to five years, which necessitates twenty to twenty-five data collections. With such a long timeframe, some establishments will inevitably discontinue providing updated benefit information after they initially participate. For many data elements, such as key provisions of benefit plans, it makes sense to hold their values constant from the last successful data collection. However, the cost for benefit plans will change over time, even if the plans' provisions do not. Therefore, an imputation procedure for the updated cost per employee is required.

The imputed cost per employee for the current quarter is found by multiplying the cost for the previous quarter by an imputed rate of change, which is given by a regression of the following form.

$$\Delta\ln C_{ijt} = \alpha + \omega\ln(W_{jt}) + \left(\sum_{k=2}^{9}\beta_k MID_{kj}\right) + \left(\sum_{k=2}^{9}\gamma_k MOG_{kj}\right) + \left(\sum_{k=2}^{4}\delta_k S_{kj}\right) + \left(\sum_{k=2}^{9}\lambda_k A_{kj}\right) + \chi FTPT_j + \phi U_j + \eta TI_j + \left(\sum_{k=1}^{N}\kappa_k KP_{kij}\right) + \varepsilon_{ijt}$$

The dependent variable is the change in the logs in the cost per participant for the plan. The set of explanatory variables is the same as for the level regression. The set of donors is trimmed for both particularly large negative and positive changes, as the top and bottom one percent of donors are excluded from the regressions. The predicted change in the logs is converted back to a ratio using the exponential transformation. The ratio is then multiplied by the cost per participant for the previous period to find the imputed cost per participant for the current period. Finally, the imputed cost per participant is multiplied by the participation rate to give the imputed cost per employee.

Imputing the rate of change through a log-change regression, then applying it to the previous value proved better in terms of the out-of-sample predictions than two obvious alternatives. It does better than ignoring the prior cost and using the level regression to impute the cost for the current quarter, even if the current quarter is several quarters removed from the last successful data collection. The log-change regression also does better than using level regressions to find imputed costs for the current and previous quarters, then using the implied rate of change to update the previous cost.

Although the NCS intends to implement the imputation procedures using the cost per participant eventually, they require the collected participation rate for the plan to match the rate used implicitly in the plan's cost per employee. The two rates do not necessarily match, however, because of historical differences between the ECI and the EBS, which the NCS has yet to reconcile. The NCS will likely use imputations based on the cost per employee in the near term.

Occasionally, the field economist will be unable to collect cost data for a given benefit area or plan, but will be able to get it for a group of benefit areas and/or plans. In these cases, the aggregate cost data must be allocated between the different benefit areas and/or plans. Allocation across benefit areas is carried out by dividing the sample from the previous quarters into cells defined by industry, occupation, ownership (i.e., private industry, state government or local government) and benefit area. Average costs are determined for each cell. Then if an aggregate cost is collected for two or more benefit areas, the average costs from the appropriate industry, occupation, ownership cell for each benefit area included in the aggregation are summed up and the percentage of that sum accounted for by each benefit area is calculated. The aggregate cost is then allocated across benefit areas using these percentages. For example, suppose short-term disability (STD) and long-term disability (LTD) costs were collected as an aggregate figure of six cents per hour worked. If, for that industry, occupation, ownership cell, STD has an average cost of 3 cents and LTD costs 2 cents, then STD will receive 60% of the aggregate cost (3.6 cents) and LTD will receive 40% (2.4 cents).

When an aggregate cost must be allocated across multiple plans within a benefit area, a different procedure is followed. First, the costs for the individual plans are imputed as if no aggregate data was available and the costs were simply missing. Then those imputations are scaled up or down proportionately so that the total matches the aggregate figure.

Finally, certain benefit areas have their costs imputed in a different manner. For overtime and leave benefits (vacation, holidays, sick leave, and other leave), costs are calculated at the occupation's average wage times the average hours of paid time off the benefit affords. For Social Security, Medicare, and federal unemployment benefits, the average gross earnings (which is the sum of wages and certain benefits) of the occupation are multiplied by the legally mandated rates (e.g., 6.2% on the first $84,900 for Social Security, 1.45% for Medicare with no upper limit, and 0.8% on the first $7000 for federal unemployment.

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

References

Barsky, C.B., Buszuwski, J.A., Ernst, L.R., Lettau, M.K., Loewenstein, M.A., Pierce, W.B., Ponikowski, C.H., Smith, J.E., and West S.A. (2000), "Alternative Imputation Models for Wage Related Data Collected from Establishment Surveys," *Proceedings of the Second International Conference on Establishment Surveys (ICES II)*, Alexandria, VA: American Statistical Association, 619-628.

Montaquila, J.M., and Ponikowski, C.H. (1993), "Comparison of Methods for Imputing Missing Responses in an Establishment Survey," *Proceedings of the Survey Research Methods Section*, Alexandria, VA: American Statistical Association, 446-451.

Montaquila, J.M., and Ponikowski, C.H. (1995), "An Evaluation of Alternative Imputation Methods," *Proceedings of the Survey Research Methods Section*, Alexandria, VA: American Statistical Association, 281-286.

Elmore, D., Loewenstein, M., and Buszuwski, J. (2001), "The Imputation of Benefit Incidence, the Number of Plans, Key Provisions, Employer Costs and Employee Costs: A Sequential Approach," *BLS internal memorandum* dated 8/08/2001.

Ernst, Lawrence R., Guciardo, Christopher J., Ponikowski, Chester H., and Tehonica, Jason (2002), "Sample Allocation and Selection for the National Compensation Survey," Proceedings of the Survey Research Methods Section, Alexandria, VA: American Statistical Association, __

Kalton, G., and Kasprzyk, D. (1982), "Imputing for Missing Survey Responses," *Proceedings of the Survey Research Methods Section*, Alexandria, VA: American Statistical Association, 22-31.