Accounting for missing data in the Employment Cost Index

Employers do not always provide all the information needed to compile the Employment Cost Index (ECI); new ECI procedures have improved the methods for dealing with missing values

Song Yi

The Employment Cost Index (ECI) is a measure of change in employer costs for employee compensation (wages and benefits). The index is compiled from information provided by employers and updated quarterly. Yet employers do not always provide all the information needed. When employers cannot provide the data, the missing values are imputed. Improvements in the method used to impute values were implemented in the March 2006 index, along with other changes—including the switch in industry classification from the Standard Industrial Classification (SIC) to the North American Industry Classification System (NAICS), and the switch from the Occupational Classification System (OCS) to the Standard Occupational Classification (SOC). The new imputation procedures incorporate some of the data available from the ongoing integration of ECI and other National Compensation Survey (NCS) products.[1]

Why are data missing?

The ECI is a voluntary survey. BLS economists contact establishments quarterly to collect compensation data for sampled occupations. Establishments are not legally mandated to respond to this request. Some may refuse to cooperate at all; others may provide partial or incomplete data. Sometimes data are available when the respondent is first contacted, but unavailable for subsequent updates.

Respondents have offered a variety of reasons why they are not able to provide data. For example, an employer may be able to provide the costs of all benefits except for a defined-benefit pension plan, the cost of which is maintained by an actuary or considered confidential. Sometimes, respondents would like to participate in the survey but are hindered by natural disasters or the absence of key staff assigned to handle the data request. Some cite the burden of multiple updates, time constraints, confidentiality issues, and the complexity of the data request.

Deriving costs with complete data

A wide array of wage and benefit information is collected from a sample of occupations within sampled establishments in selected areas to represent the civilian labor force of the entire United States.[2] Ideally, the data collected include wages and salaries, the work schedule (the number of daily and weekly hours and the number of weeks per year that employees are scheduled to work), and the costs of the following employer-provided benefits[3]:

- Paid leave—vacations, holidays, sick leave, and other leave;
- Supplemental pay—premium pay for work in addition to the regular work schedule (for example, overtime pay and pay for working on holidays and weekends), shift differentials, and nonproduction bonuses (such as lump-sum payments provided in lieu of wage increases);
- Insurance benefits—life, health, short-term disability, and long-term disability insurance;
- Retirement—defined-benefit and defined-contribution plans; and
- Legally required benefits—Social Security, Medicare, Federal and State unemployment insurance, and workers' compensation insurance.

Song Yi is a labor economist in the Office of Compensation and Working Conditions, Bureau of Labor Statistics. E-mail: yi.song@bls.gov

Respondents provide data on benefit costs in one of two forms. Preferably, data are collected on the cost (or rate) of each benefit plan and the corresponding employee participation (or usage) in each plan. If rate and usage data are not available, data may be collected for past expenditures, the amount an employer spent on a benefit for a specified time (for example, the employer's quarterly contribution to the defined-benefit pension plan).

Regardless of the format of the collected costs, the cost of each benefit is converted to a cost per hour worked. (See exhibit 1.) This allows for a consistent unit of measure to facilitate summing the components to derive a cost of total compensation (and ultimately to compare the cost from one period to the next to derive a rate of change and calculate an index). The cost per hour worked is computed by dividing the annual cost of each benefit by the annual number of hours worked, which is derived by determining the number of scheduled work hours, subtracting leave hours (such as vacation), and adding overtime hours.

In addition to being used in the calculation of the annual number of hours worked, leave and overtime hours are used in the formulas for determining the cost of these benefits. For example, typically the number of annual vacation hours is multiplied by the hourly wage rate and divided by the number of annual hours worked.

Calculations of the cost of insurance and other benefits that are not based on the wage rate may be derived from rate and usage information. For example, health insurance costs may be based on the monthly employer contribution rate for single and family coverage and the number of employees participating in each plan.[4]

How ECI deals with missing data

How missing data are dealt with depends on the extent to which they are missing—totally or partially.

Total or unit nonresponse. If at initiation (the first time data are collected from an establishment) a respondent refuses to provide information on core data elements, the situation is considered a complete refusal or unit nonresponse. Core data elements define the sampled occupation in an establishment using both establishment and occupational characteristics:

- Ownership (private industry, or State and local government)
- Industry
- Number of employees
- Geographic area
- Occupation
- Full-time/part-time status
- Collective bargaining status
- Basis for wage rates (time or incentive)

If wage data or the work schedule is not available at initiation, the situation also is considered unit nonresponse. Wages are a critical component of compensation. Wages account for about 70 percent of total compensation, and two-thirds of the benefit costs are related to wages. For example, the cost of vacations is derived by multiplying the number of vacation hours by the wage rate. Thus, about 85 percent of compensation consists of wages or is based on the wage rate.

Unit nonresponse is treated with weight adjustments that redistribute the weights of nonrespondents to similar respondents based on industry, number of employees, and geographic area. This procedure for handling refusals at initiation has not changed.

Partial or item nonresponse. If a respondent furnishes incomplete information at initiation or update, the situation is known as item nonresponse. There are three options for resolving this situation. The first is simply to ignore the missing data, a procedure that is equivalent to assuming that all missing values correspond to the absence of a plan, in which case the cost of the benefit is 0. This option by definition understates the average cost in the aggregate. The second option is to generate some random number to assign a value. Neither of these alternatives is credible. The third approach, which is used for ECI, imputes an estimate for the missing data using information obtained from other similar establishments and employees. The imputed value provides a best guess for the missing data.

Imputation involves matching donors (which have the information) with recipients (which do not). Approaches for imputing estimates for missing values are described in the box on page 25.

Change in imputation procedures

As of March 2006, BLS changed the procedures used to impute missing values. The change incorporates the switch in industry and occupational classification from SIC-OCS to NAICS-SOC—two of the core data elements used to match donors and recipients.[5]

The following discussion highlights the differences between the old and new imputation methods. Different procedures are used at initiation and for quarterly updates, as well as for estimating missing wages, hours, and benefit costs. The new procedures for estimating missing hours and benefit costs have replaced the cell-means method with a combined nearest neighbor and regression approach and slightly modified the regression formula used to update missing wages. (See exhibit 2.)

Wages. The procedure for handling missing wage data is basically unchanged. As previously noted, if wage data are unavailable at initiation, the establishment is considered a total nonresponse. If wage data were collected at initiation, but not at a quarterly update, a rate of wage change is imputed. The

change rate is applied to the previously collected wage data to calculate a new imputed wage rate, which is then used in subsequent calculations of the costs of wage-related benefits. However, there is a slight change in the specification of the regression equation used to estimate the rate of change.[6]

Hours. Imputation of missing hours occurs only during the initial collection of data. While the number of hours is held constant at the quarterly updates, the cost of the hours-related benefits is updated to reflect changes in the wage rate. There is no change in this practice. However, the procedure used to impute missing hours at initiation has changed. Data are imputed for missing hours for overtime, vacation, sick leave, holidays, other paid leave, and unpaid leave.

Under the new procedure, the imputed hours are used for the calculation of both the number of annual hours worked and the corresponding cost of the hours-based benefit. The cost of vacation then is derived by multiplying the wage rate by the number of imputed vacation hours. Thus, with the new procedure, in calculating annual hours and the cost of the

Exhibit 1. Sample ECI calculations

Annual hours worked

Item	Schedule/benefit	Calculation	Hours
Scheduled work hours ..	8 hours per day		8
	5 days per week	5 × 8 =	40
	52 weeks per year	52 × 40 =	2,080
Leave:			
Vacation	3 weeks	3 × 40 =	120
Holidays	8 days	8 × 8 =	64
Sick leave	3 days	3 × 8 =	24
Overtime	50 hours		50

Annual hours worked = Scheduled work hours - leave hours + overtime hours = 2,080 - (120 + 64 + 24) + 50 =1,922 (1)

Cost per hour worked for vacations

Hourly rate	$10.00
Vacation leave hours	120

Cost per hour worked for vacations = (Hourly rate × leave hours) / annual hours worked = ($10 × 120) / 1,922 = $0.62 (2)

Cost per hour worked for health insurance based on rate and usage

Health coverage type	Monthly cost (rate)	Employees enrolled in plan (usage)
Family	$300	1
Single	$200	1
None	$0	1

Cost per hour worked for health insurance = [(Rate × number with family coverage) + (rate × number with single coverage) + (rate × number not covered)] × 12 months / annual hours worked = [(1 × $300) + (1 × $200) + (1 × 0)] × 12 / 1,922 = $3.12 (3)

Cost per hour worked for defined-benefit retirement plans based on expenditure data

Period	Expenditure on plans per employee
Quarter	$100
Year	$1,200

Cost per hour worked for defined-benefit retirement plans = Annual expenditure / annual hours worked = $1,200 / 1,922 = $0.62 (4)

Imputation methods

Nearest neighbor. Imputation classes (cells) are formed based on auxiliary data that are known for all units, such as ownership (private industry, or State or local government), industry, major occupational group, an indicator of whether the benefit cost is wage-related, collective bargaining status, region, and full- or part-time status. Within each cell, a unit that is missing the characteristic of interest (known as a recipient unit) takes the value of the characteristic of a usable unit (known as a donor unit) that is nearest to the recipient. "Nearest" is defined by the similarity of auxiliary data available for donors and recipients. There may be many potential donors from the same cell, so the decision as to which donor the recipient will use is based on establishment size; the donor with establishment size closest to the recipient's is chosen for imputation. Still, there may be cases in which donors do not have all of the same characteristics as the recipient. In this case, the nearest neighbor procedure drops one of the characteristic variables to see whether a full match then can be obtained. If no full match can be obtained even after one variable is dropped, another variable is dropped until a full match can be found.

Cell means. Imputation cells are formed in much the same way as for the nearest neighbor method. However, instead of using the characteristics of interest from a single donor in the cell for the imputation, the mean of all donors in the cell is calculated. Then the mean value is imputed to the recipient. To ensure that the number of donors in each cell is fairly large, cells generally are defined by industry and major occupation group only.

Regression. Imputations through regression models are similar to imputations through the cell-means method. They both use values for the characteristics of interest from multiple donors rather than from a single donor. However, instead of using the average value among donors from a particular cell, the regression imputation uses the data from all donor cases to arrive at the best predicted value. The regression equation is used to impute the missing characteristic of interest, based on the recipient's values for the variables on the right-hand side of the regression equation:

$$Y = a + b_1X_1 + b_2X_2 + ... + b_pX_p$$

where the coefficients for the right-hand side variables include ownership (private industry, or State or local government), industry group, major occupational group, collective bargaining status, full- or part-time status, locality (area), and number of employees. The explanatory variables on the right-hand side differ according to the specific variable that is imputed.

benefit, the number of imputed hours is used as if actual hours had been collected. For example, if imputed hours are handled as collected vacation hours in exhibit 1 (equations 1 and 2), the imputed number of vacation hours (120 hours) is used not only in the calculation of the annual number of hours worked, but also in the calculation of the cost of vacations ($0.62 per hour worked).

This is a departure from the old method, according to which the number of hours and the corresponding costs were imputed independently. Previously, there was no attempt to link the number of vacation hours with the corresponding cost of the vacation benefit. These independent computations might yield inconsistent results.

The new method for imputing the cost of the hours-based benefits is a three-step process. First, when it is not known whether a leave or overtime plan exists, a nearest neighbor imputation procedure is used to make this determination. This first step of the process is new; it uses information available from the integration of ECI and other NCS products. If it is imputed that there is no plan, there is no cost. If, however, a plan exists, a regression formula is used to impute the number of hours and subsequently the cost. The explanatory variables for a given hours-based benefit (for example, sick leave) include all the core data elements (also used in the wage imputation equation), as well as the number of hours for the other hours-based benefits. Thus, the explanatory variables used in the regression formula for missing paid sick leave include vacation, holiday, and other paid leave.

In the past, the number of missing hours and the cost of the hours-based benefits were imputed using a cell-means method. The cell-means approach aggregates a large sample of donors with similar characteristics and obtains an average for the missing hours worked. For more information, see the box above.

Benefit costs. The benefit imputation procedure is used to estimate missing employer costs for health insurance, life insurance, long-term and short-term disability insurance, retirement (defined-benefit and defined-contribution) plans, shift differentials, nonproduction bonuses, State unemployment insurance, and workers' compensation insurance. Social Security, Medicare, and Federal Unemployment Insurance are not imputed, because these costs can be derived.[7] The procedure used to impute missing benefit costs has changed both at initiation and for quarterly updates.

Exhibit 2. Comparison of ECI imputation methods[1]

Method	Missing information: Wages	Missing information: Number of hours and corresponding costs[2]	Missing information: Other benefits[3]
New:			
Initiation	The unit continues to be treated as a total refusal, and the assigned sample weight is reallocated to other units.	1. The nearest-neighbor approach is used to determine if there is a plan. 2. Regression is used to impute hours of individual hours-related benefits, which then are used to compute the annual hours worked. 3. The imputed number of hours is used to estimate the cost of the corresponding benefit.	1. The nearest-neighbor approach is used to determine if there is a plan. 2. Regression is used to impute missing costs.
Update	A minor change to the regression is used to impute wage change.	Hours are held constant at update; cost is updated to reflect changes in wage rates.	Regression is used to impute the rate of change.
Old:			
Initiation	The unit was treated as a total refusal; the assigned sample weight was reallocated to other units.	A cell-means approach was used to impute hours to derive annual hours worked, and a separate cell-means imputation was done to impute the cost of the benefit.	A cell-means approach was used to impute benefit cost rates.
Update	Regression was used to impute wage change.	Hours were held constant at update; cost was updated to reflect changes in wage rates.	A cell-means approach was used to impute the rate of change.

[1] In all cases the North American Industry Classification System (NAICS) has replaced the Standard Industrial Classification (SIC), and the Standard Occupational Classification (SOC) has replaced the Occupational Classification System (OCS).

[2] Hours are imputed for overtime, vacation, holidays, sick leave, other paid leave, and unpaid leave.

[3] Benefit costs are imputed for health insurance, life insurance, retirement (defined-benefit plans and defined-contribution plans), long-term and short-term disability, shift differentials, nonproduction bonuses, State unemployment insurance, and workers' compensation insurance. Social Security, Medicare, and Federal unemployment insurance are not imputed.

At initiation, there is a two-step process similar to the one used for imputing the number of leave and overtime hours. First, if it is not known whether the establishment offers a benefit, a nearest neighbor approach is used to make this determination. If it is imputed that there is no plan, there is no cost. Otherwise, a regression equation is used to impute the missing costs. The explanatory variables in the formula are the core data elements and the wage rate.

This process differs from the past. Previously, a cell-means method was used to estimate the missing costs, and there was no procedure for first determining if a plan existed. The integration of ECI and other NCS products provides the additional information to make the new process viable.

For quarterly updates, the new procedure imputes a rate of change for missing benefit costs using a regression. The explanatory variables are the same as those used for the regression imputation at initiation. This procedure replaces the cell-means approach to imputing a rate of change.

There also is a change in the procedure used for handling multiple quarters of missing data. In the past, if wage and benefit data were not updated for more than four consecutive quarters, the establishment was considered a refusal. Under the new pro-

cedure, missing data are imputed as long as the establishment remains in the sample.

Allocation. Although data on the cost of individual benefits are always preferred, in some cases an employer can provide data only for combined benefits. For example, an employer may report the costs of all the insurance items combined or "collapsed" because the insurance benefits were obtained as a package, and the bill does not separate the components. In such situations, the collapsed cost is allocated to the individual benefits based on the proportional costs of those benefits among establishments and occupations with similar characteristics.

For example, the cost of insurance (consisting of health and life insurance) for a given occupation is $5 per hour worked, and the average collected costs for occupations with matching industry and occupational characteristics is $2 per hour worked for health insurance and $.50 per hour worked for life insurance (80 percent and 20 percent, respectively, of the total insurance cost). These percentages are allocated to the $5 per hour worked combined health and life insurance cost—or $4 and $1, respectively.

This allocation method has remained basically unchanged. But, as with all imputation procedures, the allocation process reflects the change in industry and occupational classifications.

Impact on published estimates

Preliminary research suggests that the changes to the imputation procedures will have a relatively small impact on the vast majority of the published ECI and Employer Cost for Employee Compensation (ECEC) series.[8] Both of these series are components of NCS and use the same data. ECI shows the rate of change in compensation costs, while ECEC presents the compensation costs in terms of cost per hour worked. The new imputation procedures tend to increase the average cost for the hours-based benefits (paid leave and overtime) by a few cents, while tending to decrease the average cost for the other benefit areas, such as health insurance and retirement plans, by a few cents. The net result typically is small, particularly for the closely watched aggregate estimates of total compensation for all civilian and all private industry workers.

However, the new imputation procedures have a more substantial impact on ECEC estimates for a few disaggregate groups, such as union workers. This effect occurs when the characteristic that defines the particular group has been used as a variable to match observations with missing data to observations that reported data. Two examples are the ECEC estimate for union workers and the ECEC estimate for part-time workers. The previous imputation procedures did not use these characteristics in matching donors and recipients. Therefore, the compensation for some nonunion workers contributed to the ECEC estimate for union workers. All else equal, this tended to lower the ECEC estimate for union workers. Similarly, the compensation for some full-time workers contributed to the ECEC estimate for part-time workers, which tended to raise the ECEC estimate for part-time workers.

Uses of integrated NCS data

NCS collects information on the provisions and costs for an extensive set of benefits, including health insurance and retirement plans. Recent efforts have focused on collecting the data for all NCS products in one vehicle. The integration of data on the provisions of benefit plans with data on costs allows statisticians to calculate costs associated with a particular provision, such as a prepaid funding arrangement for a health insurance plan. Such characteristics of a benefit plan also may be used to improve the imputed cost for the benefit when the characteristics are available but the cost is missing. □

Notes

[1] The National Compensation Survey (NCS) provides measures of employer costs for wages, salaries, and benefits, as well as details of employer-provided benefits. NCS continues to publish data series previously produced by the three BLS programs it replaced: (1) the Occupational Compensation Survey, (2) the Employee Benefits Survey, and (3) the Employment Cost Index. The integration of the sample selection, data collection, and processing for these programs allows for data obtained for one series to be used in compiling data for others.

[2] For a description of the procedure used to develop the area, establishment, and occupational sample, see chapter 8 of the *Handbook of Methods*, www.bls.gov/opub/hom/pdf/homch8.pdf.

[3] Prior to March 2006, the ECI definition of compensation included the cost of severance pay and Supplemental Unemployment Benefits (SUB) plans. A discussion of the reasons for dropping these benefits appears in Fehmida Sleemi, "Employment Cost Index publication plans," in this issue.

[4] For a more detailed discussion of the method for computing the cost per hour worked, see John Ruser, "The Employment Cost Index: what is it?" *Monthly Labor Review,* September 2001, pp. 3–16.

[5] See Sleemi, "Employment Cost Index publication plans."

[6] In the new method, the dependent variable is the natural log of the ratio of the current average hourly rate to the prior quarter's average hourly rate. Previously, the dependent variable used the ratio of the current-quarter hourly rate to the prior-quarter average hourly rate. The explanatory variables on the right-hand side of the equation consist of the core data elements.

[7] The cost of Social Security, Medicare, and Workers' Compensation is derived by multiplying gross annual earnings (annual wages, overtime, paid leave, shift differentials, and nonproduction bonuses) by published rates. Employers are required to contribute to these benefits.

[8] Because the same data are used in compiling the ECI and the ECEC, imputed values for the ECI also are used in ECEC tabulations.