ECI SURVEY DESIGN: AN UPDATE

Steven Kaufman, Bureau of Labor Statistics

Introduction

The Employment Cost Index (ECI) is a relatively new Bureau of Labor Statistics survey measuring the change in the employer cost of employing workers. When the ECI first started its publication in December 1975, it measured quarterly wage change covering the private non-farm sector excluding Alaska, Hawaii and private households. Publications included overall National, Major Industry Division (MID), Major Occupation Group (MOG), Census Region, Union/Non Union and Metropolitan/Non-Metropolitan Area quarterly change numbers. Currently, the ECI is an index measuring total compensation change covering the total non-farm civilian sector excluding private household and the federal government. Compensation is composed of wages and twenty-three benefits. The National series (Overall National, MID, MOG indices) use Laspeyres (fixed weight) estimates. The Non-national series (region, union and metropolitan indices) use current weights to compute indices.

As a new and evolving survey, the ECI has encountered problems. The original estimation system produced independent National and Non-national estimates. As long as the sample remained fixed, all estimates were internally consistent. As the sample changed, inconsistencies between the National and Non-national estimates became apparent. A new estimation system, introduced in June 1981, has been developed to correct this inconsistency problem. This is the first topic of the paper.

A second problem concerned the establishment frame. The original survey design [1/] did not incorporate procedures to update the frame. A periodic survey like the ECI must develop methodology to update the frame. Otherwise, as the frame ages, current estimates can no longer be considered true current estimates. Our solution has been to replace (replenish) a portion of the sample each quarter. This is the second topic of the paper.

Estimation

Original (1975) ECI Estimation System [2/]

ECI produces two types of quarterly change series, National and Non-national. The overall National, MOG, and MID are National series; and the Regional (4-Census Regions), Union (union/nonunion) and Metropolitan (Metropolitan/Non Metropolitan) are the Non-national series. In the original estimation system, a national estimate, $R^t_{N_i}$, had the following form:

$$R^t_{N_i} = \frac{\sum_{kj \varepsilon N_i} C_{kj} \; X^t_{kj}}{\sum_{kj \varepsilon N_i} C_{kj} \; X^{t-1}_{kj}},$$

where:

C_{kj} is the 1970 Census Weight for k^{th} 2 digit SIC (Standard Industrial Classification), j^{th} Census Occupation;

$\bar{X}^t_{kj}$ is the sampled weighted average wage for the kj^{th} cell at time t, computed with matched quotes [3/] at both time periods t and t-1;

$\bar{X}^{t-1}_{kj}$ is the sampled weighted average wage for the kj^{th} cell at time t-1, computed with matched quotes at both time periods t and t-1; and

N_i is either the overall National, an MOG, or an MID series i.

A Non-national quarterly change series, $R^t_{S_{il}}$, had the following form:

$$R^t_{S_{il}} = \frac{{}_{kj \quad S_{il}} \; C_{kj} \; P^o_{kjS_{il}} \; \bar{X}^t_{kjS_{il}}}{{}_{kj \quad S_{il}} \; C_{kj} \; P^o_{kjS_{il}} \; \bar{X}^{t-1}_{kjS_{il}}},$$

where:

C_{kj} is the 1970 Census Weight for the k^{th} 2 digit SIC, j^{th} occupation;

$P^o_{kjS_{il}}$ is the proportion of the sampled weights for the l^{th} component of the S_{il} Non-national series (union, metropolitan or region series) within the kj^{th} cell. This proportion is calculated using the December 1975 data;

$\bar{X}^t_{kjS_{il}}$ is the sampled weighted average wage for the l^{th} component of S_{il} Non-national series within the kj^{th} cell at time t, computed with matched quotes at both periods t and t-1; and

$\bar{X}^{t-1}_{kjS_{il}}$ is the sampled weighted average wage for the l^{th} component of the S_{il} Non-national series within the kj^{th} cell at time t-1, computed with matched quotes at both time periods t and t-1.

Consistency

The original ECI estimation system produced inconsistent estimates. On occasion, the value of national estimate did not lie between the values of the non-national estimates (e.g., the metro and non-metro estimates between December 1976 and December 1977 were 6.9%, and 6.7% respectively while the national estimate was 7%). To ensure this type of inconsistency does not occur, it is necessary that the numerator (denominator) of the national estimate can be produced by summing the appropriate non-national numerators (denominators). If this property holds, the estimates will be consistent. That is, consistent estimates can not be calculated independently; instead, they must build on each other from lower level estimates that cross all publication criteria. Once this is accomplished, all estimates can be obtained by summing these lower level estimates.

There were two major reasons for the original system's inconsistent estimates. The first reason came from the non-national employment estimates $(C_{kj}P^{o}_{kjS_{il}})$ being fixed over time. As the responding sample changes, either by increased refusals or additions or replacements to the current sample, the set of non-national cells directly covered by responding units also changes. If there were no respondents in a cell when the $P^{o}_{kjS_{il}}$ were calculated, any responses added in the cell since time 0 were not used in the particular non-national estimate. However, these data were used in the national estimates because these estimates were calculated independently without using the $P^{o}_{kjS_{il}}$.

The second reason for inconsistencies arose when change figures were computed for periods longer than a quarter by compounding the appropriate quarterly relatives. A typical estimate for quarters t-2 to t, $R^{t}_{(t-2)i}$ was:

$$R^{t}_{(t-2)i} = R^{t}_{i} \times R^{t-1}_{i} = \frac{X_{ti}}{X_{(t-1)i}} \times \frac{\hat{X}_{(t-1)i}}{X_{(t-2)i}}$$

where:

X_{ti} and $X_{(t-1)i}$ are the numerator and denominator of the quarter t relative, and

$\hat{X}_{(t-1)i}$ and $X_{(t-2)i}$ are the numerator and denominator of the quarter t-1 relative.

$X_{(t-1)i}$ and $\hat{X}_{(t-1)i}$ both estimated the cost at time t-1. Ideally, they would be equal, and $R^{t}_{(t-2)i}$ would be the employer cost for quarter t divided by the employer cost for quarter t-2. If these costs were computed by summing costs at sic/occ/reg/union/metro level, all estimates would be consistent. However, if $X_{(t-1)i}$ and $\hat{X}_{(t-1)i}$ were not equal, then consistent estimates would require the sum of $X_{ti}\ \hat{X}_{(t-1)i}$ across appropriate subseries (i) be equal to $X_{t}\ X_{t-1}$(the product of the national estimates). Since this condition was not true in general, unequal t-1 cost estimates would produce inconsistent change estimates.

In the original ECI estimation system, $X_{(t-1)i}$ was based on matched quotes from quarters t-1 and t, while $\hat{X}_{(t-1)i}$ was based on matched quotes from quarters t-2 and t-1. As long as the sample remained constant, $X_{(t-1)i}$ and $\hat{X}_{(t-1)i}$ would be equal. However, during quarter t, it would be possible for a responding establishment to refuse or for new respondents to be added. In this case $X_{(t-1)i}$ and $\hat{X}_{(t-1)i}$ would not be equal. Therefore, the ECI estimates were inconsistent.

Because of the problems discussed above, the ECI using the original estimation system, can not be considered a Laspeyres (fixed weight) Index. The employment distribution (fixed $P^{o}_{kjS_{il}}$) can not be kept fixed without compromising consistency. The national estimates would be Laspeyres, if it weren't for the sic/occ cells that were periodically no longer covered because of attrition of the sample. For this reason, no indices were published until the new system, discussed below, was in place.

Laspeyres Index

A consistent Laspeyres wage (compensation) Index , I^{t}_{o}, can be defined in the following manner:

$$I^{t}_{o} = \frac{\sum_{k} CW^{o}_{k}\ M^{t}_{k}}{\sum_{k} CW^{o}_{k}},$$

where:

CW^{o}_{k} is the base period total wage cost weight (product of the Census Weight and the base period weighted average wage) for the k^{th} estimation cell;

M^{t}_{k} is the wage change in the k^{th} cell from time o (base period) to time t,

$$M^{t}_{k} = \prod_{n=1}^{t} \frac{CB^{n}_{k}}{CB^{n-1}_{k}};$$

CB^{n}_{k} is the cost bill (total wage or compensation cost) for the k^{th} cell during time n using matched quotes from times n and n-1; and

CB^{n-1}_{k} is the cost bill for the k^{th} cell during time n-1 using matched quotes from times n and n-1.

This is a Laspeyres estimator, in that the base period cost weights are fixed and are used to weight the wage change (M^{t}_{k}) across the k cells. It will be consistent provided all subindices are computed using only subsets of the k cells.

An ECI Laspeyres Index based on 23 occupational estimation cells within each SIC requires 22,816 cells (62 SIC x 23 occupation x 4 region x 2 union x 2 metro). The ECI sample size of 10,000 quotes can not support all these cells. Therefore, compromises must be made between sample resources and the desire for consistent, Laspeyres ECI estimates.

Estimation Options

One approach is to use the above formula and reduce the number of estimation cells. This would provide Laspeyres estimates, but the quality and/or number of published estimates must suffer. The overall National, Industrial and MOG series are the most important. Publishing only these national series and eliminating publication of the region, union and metropolitan series requires 1426 (62 x 23) estimation cells. Our sample would support these cells and provide Laspeyres National series. However, region, union and metro series are deemed too important to stop publication.

To avoid eliminating the non-national series, it would be possible to combine 2 digit SICs to reduce the number of cells. However, wages and compensation movement can vary greatly across SICs. Combining such cells would increase the within cell variance and hurt our reliability. In addition, ECI estimates, to the extent possible, should not measure change caused from employment shifts, which would not be controllable within collapsed cells. Therefore, collapsing SICs is not acceptable.

Union, region and metro cells could be combined based on similar rates of change. If such cells are determined and held fixed, then a meaningful Laspeyres index can be produced. However, over time one would expect that the union, region and metro relationships would change. To maintain a meaningful index, the estimation cells accordingly should be changed. Once the cells are modified, the internal weighting structure would be different, an unacceptable situation for a Laspeyres index.

We concluded a Laspeyres index, which we would want to publish, can not be produced without reducing the number of published estimates. We decided to introduce a new methodology, a variation of the Laspeyres index above to produce current ECI estimates. Now, all National series (Industrial and Major Occupation Group) use consistent and Laspeyres estimators. Each non-national quarterly estimate is consistent with the national estimate. Non-national estimates for periods longer than a quarter are generally not consistent. None of the non-national estimates use Laspeyres estimators.

Current ECI Estimator

The National estimates have the following form:

$$I_{N_i} = \frac{\sum\limits_{kjc\,\varepsilon\,N_i} CW^o_{kjc}\; M^t_{kjc}}{\sum\limits_{kjc\,\varepsilon\,N_i} CW^o_{kjc}} = \frac{\sum\limits_{kjc\,\varepsilon\,N_i} CW^t_{kjc}}{\sum\limits_{kjc\,\varepsilon\,N_i} CW^o_{kjc}},$$

where:

c indicates whether a cost weight is for wages or total benefits;

CW^o_{kjc} is the base period cth type cost weight for the kth 2 digit SIC, jth Census occupaton; and

M^t_{kjc} is the change in the cth type of cost within the kjth cell from times o to t

$$M^t_{kjc} = \prod_{n=1}^{t} \frac{CB^n_{kjc}}{CB^{n-1}_{kjc}},$$

where:

CB^n_{kjc} is the cost bill for the cth type of cost within the kjth cell during quarter n using matched quotes from quarters n and n-1;

CB^{n-1}_{kjc} is the cost bill for the cth type of cost within the kjth cell during quarter n-1 using matched quotes from quarter n and n-1; and

CW^t_{kjc} is the quarter t cost weight for the kjcth cell.

This is the Laspeyres index defined above using SIC and occupation cells. There are wage, total benefit and total compensation (sum of wage and total benefits) cost weights. This means that any estimate that can be produced by summing these variables, all ECI national estimates, will be both Laspeyres and consistent.

Rate of Change from time t to T ($R^T_{tN_i}$) is:

$$R^T_{tN_i} = \frac{I^T_{N_i}}{I^t_{N_i}} = \frac{\sum\limits_{kjc\varepsilon N_i} CW^T_{kjc}}{\sum\limits_{kjc\varepsilon N_i} CW^t_{kjc}}.$$

Again, all estimates, including all national series, produced by summing the kjc cells will be both Laspeyres and consistent.

The Non-national quarterly estimates[4] are computed independently for each type of Non-national series (region, union and metropolitan series) using:

$$R^t_{S_{il}} = \frac{\sum\limits_{kjc\varepsilon S_{il}} CW^{t-1}_{kjc}\; q^{t-1}_{kjcS_{il}}\; r^t_{kjcS_{il}}}{\sum\limits_{kjc\varepsilon S_{il}} CW^{t-1}_{kjc}\; q^{t-1}_{kjcS_{il}}}$$

where:

c represents a type of measured cost, either wages or total benefits;

CW^{t-1}_{kjc} is the cth type of cost weight within the kth 2 digit SIC, jth census occupation during quarter t-1;

$q^{t-1}_{kjcS_{il}}$ is the proportion of the S_{ij} subseries cost bill within the kjth cell using quarter t-1 data matched from quarters t-1 and t ($CB^{t-1}_{kjcS_{il}} / CB^{t-1}_{kjc}$); and

$r^{t}_{kjcS_{il}}$ is the quarterly change during quarter t in the cth type of cost bill for the S_{il} non-national subseries within the kjth cell using only matched quotes from quarters t-1 and t ($CB^{t}_{kjcS_{il}}$ / $CB^{t-1}_{kjcS_{il}}$).

Now, for quarter t,

$$\sum_{l} CW^{t-1}_{kjc}\; q^{t-1}_{kjcS_{il}}\; r^{t}_{kjcS_{il}}$$

$$= \sum_{l} CW^{t-1}_{kjc}\; \frac{CB^{t-1}_{kjcS_{il}}}{CB^{t-1}_{kjc}}\; \frac{CB^{t}_{kjcS_{il}}}{CB^{t-1}_{kjcS_{il}}}$$

$$= CW^{t-1}_{kjc}\; {}_{l}\; \frac{CB^{t}_{kjcS_{il}}}{CB^{t-1}_{kjc}}$$

$$= CW^{t-1}_{kjc}\; \frac{CB^{t}_{kjc}}{CB^{t-1}_{kjc}}$$

$$= CW^{o}_{kjc}\; M^{t}_{kjc}$$

and

$$\sum_{l} CW^{t-1}_{kjc}\; q^{t-1}_{kjcS_{il}} = \sum_{l} CW^{t-1}_{kjc}\; \frac{CB^{t-1}_{kjcS_{il}}}{CB^{t-1}_{kjc}}$$

$$= CW^{t-1}_{kjc} \sum_{l} \frac{CB^{t-1}_{kjcS_{il}}}{CB^{t-1}_{kjc}}$$

$$= CW^{t-1}_{kjc}\,.$$

Since the sum of the non-national estimates' numerators (denominators) equals the national estimate's numerator (denominator) within an sic/occupation cell, the quarterly non-national estimates are each consistent with the national estimate.

Also,

$$CW^{t-1}_{kjc}\; q^{t-1}_{kjcS_{il}} = CW^{o}_{kjc}\; M^{t-1}_{kjc}\; q^{t-1}_{kjcS_{il}}.$$

That is, the non-national cost weight is computed using a quarter t-1 distribution to split out the cost weight from the quarter t-1 national cost weight. This means that none of the non-national subseries are Laspeyres estimators.

Non-national estimates for periods longer than a quarter are formed by compounding quarterly relatives:

$$R^{t}_{(t-2)S_{il}} = R^{t}_{S_{il}} \times R^{t-1}_{S_{il}} = \frac{X^{t}_{S_{il}}}{X^{t-1}_{S_{il}}} \times \frac{\hat{X}^{t-1}_{S_{il}}}{X^{t-2}_{S_{il}}}.$$

As with the original system, two estimates for the same time period ($X^{t-1}_{S_{il}}$ and $\hat{X}^{t-1}_{S_{il}}$) are computed, based on different quarter to quarter matched quotes. This can make the two estimates different which can cause inconsistencies. However, with the new system, this situation is not as serious as the original system. The two major causes of inconsistency (loss of coverage for SIC/occ/S_{il} or SIC/occ cells) can not happen with the new system. SIC/occ/S_{il} census weights are reapportioned each quarter and an imputation is used whenever no data exists for a SIC/occ cell. Therefore, even though non-national estimates for periods longer than a quarter are inconsistent, the distortion is less than with the original system.

Replenishments

Original Survey Design[1]

The original ECI sample was selected from the 1972 Unemployment Insurance File using a two stage selection process. The first stage, a mail survey of 10,000 establishments, collected occupational employment data for approximately twenty-three SIC specific occupations. The second stage was an establishment selection of 2,000 establishments employing a two-way controlled selection controlling the respondent burden and the number of designated quotes within each selected occupation.

Since December 1975, a few supplements have been added to compensate where the original design provided inadequate coverage for a desired publication cell. Otherwise, the original sample had not changed since 1975, causing a 82% wage response rate (in 1975) to drop to 63% (in 1981). Shortly after the State and Local Government sector was added to the ECI scope in September 1981, the first replenishment sample was introduced into the ECI sample. The planning had started in late 1979, when the response rate was 67% with 1800 establishments.

Purpose of Replenishment Samples

The purpose of the ECI replenishment samples is to quarterly replace a different portion of the original private sector sample, until the entire sample is replaced in a number of years. The quarterly replenishment groups should each have an equal number of establishments. This minimizes the disruption in the quarterly estimates and is within our resource constraints. To introduce one group each quarter, a replenishment group collection cycle begins every three months. Once the six month collection cycle is completed, the first quarterly update will be analyzed for errors and compared with the original sample. Assuming all errors have been corrected, the new sample is introduced into the ECI estimates after the second update.

In terms of the survey design there are three advantages of the replenishment process. The first advantage is providing estimates based on a reasonably current frame. Upon completion of a replenishment cycle, say three years long, the first replenishment sample will be five years old, while the last replenishment sample will be two years old. Thus, the overall national, regional, union and metropolitan estimates are based on current compensation costs from establishments that represent the frame as it looked "on average" almost four years ago. Since replenishment samples are introduced by SIC (see description of sample selection procedures below), the MID estimates will be based on establishments that represent a frame 2 to 5 years old, depending on the MID.

The second advantage is a stable sample size. With a stable sample, the 1.5 to 2% quarterly sample loss from increased refusals and out of business establishments, is balanced with incoming replenishment samples. The sampling error will be more stable and the likelihood of having estimation cells with no active respondents will be reduced. By making assumptions about the out of scope, initial refusal and ongoing loss rates, the sample size can be approximated using the formula described in the next section.

The third advantage is maintaining a stable response rate. Bias is introduced from non-responding establishments (both initial and ongoing refusals). All BLS surveys have some nonresponse and assume that non-respondents are similar to respondents for estimation purposes. With a high response rate, estimates should be reasonably representative of the frame, resulting in a low nonresponse bias. A low response rate produces estimates that may not be representative of the frame, especially if the assumptions made about similar movement for nonrespondents and respondents are false. This would produce a high nonresponse bias. Therefore, the response rate can be considered an indirect indicator of nonresponse bias and a stable response rate should provide some confidence that the nonresponse bias is stable.

Determination of Maintainable Response Rate and Quarterly Sample Size

If we assume the sample is completely replaced after n, 2n, 3n... quarters and that the response and attrition rates are equal across replenishments, then the response rate obtained after n quarters should be maintained each quarter thereafter. We call this the maintainable response rate. The determination of the appropriate time length for the complete replenishment cycle can be made by computing the maintainable response rates for various cycles and comparing the rates.

To compute the maintainable response rate, the following wage information from the original sample is used:

p =	proportion of initial sample in scope	0.85
q =	proportion of initial in scope sample responding	0.82
r =	proportion of sample remaining each quarter	0.98
m =	number of establishments required at end of replenishment cycle	2000

The benefit response rate is approximately 10 percentage points lower than the wage response rate.

Using a little algebra the maintainable response rate MR is:

$$MR = \frac{q(1-r^L)}{L(1-r)} ,$$

where L is the number of quarters required to complete the replenishment cycle.

The number of units to be introduced each quarter (N) is

$$N = \frac{m(1-r)}{(1-r^L)pq} .$$

From these two formulas and the figures provided, the following table of quarterly sample size and maintainable response rates can be produced.

Replenishment Cycle (Year)	Number of units initiated each quarter	Maintainable response rate (wages)
2	385	0.76
3	267	0.74
4	208	0.71

Considering the initial work required introducing an establishment into the survey, a two year cycle was not considered cost effective. A 0.71 wage response rate with a four year cycle is rather low considering the fact that the benefit response rate would be closer to 0.6 than to 0.7. A three year cycle would keep respondents in the survey for a reasonable length of time and provide a benefit response rate at least close to 0.65. Therefore, the initial decision was to proceed with a three year cycle. After the first year of replenishment samples, it became apparent that field resource constraints would not allow a three year cycle. We are currently working on a four year cycle.

Description of the Establishment Selection

The replenishment samples are introduced each quarter stratified by two digit SIC. A set of related SICs is chosen each quarter. Within each SIC, the frame may be sorted by Census Region, employment or alphabetical order. Within each SIC, a probability proportionate to employment sample is selected with approximately 400 establishments for the entire replenishment group. Systematic samples of about 300 establishments comprise the main replenishment sample. The remaining 100 establishments are systematically placed into several supplemental groups. The supplemental groups are held in reserve in case additional sample is required. This usually occurs if a larger than expected number of out of scope establishments is observed. Finally, to help variance estimation, the establishments are placed into two half-samples.

Description of the Occupation Selection

In order to measure MOG compensation change, the set of ECI occupations must be determined so that they represent an Occupational Universe (currently based on 1970 Census Occupation list). The original sample design used a probability selection process to select a very detailed set of Census Occupations before final collection. Even though the phase 1 occupation survey produced an occupational distribution of the detailed occupations, the number of occupations finally collected was far less than expected. This was probably caused by phase 1 occupation misclassifications, the age of the phase 1 sample, and phase 1 imputation process. To avoid the phase 1 problems and increase the number collected occupations within an establishment, a set of larger based occupations called Entry Level Occupations (ELO) have been introduced into the survey design.

An ELO is a broad SIC specific grouping of Census Occupations within an MOG. There are usually 9 to 13 ELOs, which represent all occupations within the SIC. Each ELO found in the establishment is collected. During the initiation, each detailed establishment occupation is matched into one of the ELOs. Then a probability proportionate to employment selection is made within each ELO, selecting one specific occupation. Data for wages and benefits is then collected for each of the selected detailed establishment occupations.

In summary, the ECI has progressed significantly during the past few years. The original estimation system has been replaced with a new system, providing consistent Laspeyres national estimates and consistent non-national quarterly change numbers. A new sample design has provided the ability to update the frame on a ongoing basis. Once the replenishment cycle is completed, the frame will be 'on average' four years old. As the replenishment samples are continued, the average age of the frame will be maintained.

FOOTNOTES

1/ General Survey Design Aspects of the ECI, C. Easley Hoy, 1978 ASA Proceeding, San Diego, CA.

2/ ECI Estimation Procedures, Douglas Wright and Steven Kaufman, 1978 Proceeding, San Diego, CA.

3/ A matched quote is establishment/occupation data from an establishment that has reported for the occupation in both time periods.

4/ For an economic interpretation of the non-national estimates see:
Estimation Procedures for the Employment Cost Index, G. Donald Wood, Jr., Monthly Labor Review, May 1982.

ACKNOWLEDGMENT

I would like to acknowledge the work Don Wood and Dan Kohler did on the current estimation system.

I would like to express my thanks to Jo-Ann McCrae who typed the final paper.