CE Source Selection for Publication Tables

BRETT J. CREECH AND
BARRY P. STEINBERG

The Consumer Expenditure Survey (CE) is a nationwide household survey conducted by the U.S. Bureau of Labor Statistics (BLS) to find out how Americans spend their money. The CE consists of two components, each with its own questionnaire and sample: the Interview Survey and the Diary Survey. The data are collected for the BLS by the U.S. Census Bureau. There is some overlap in the information collected by the two surveys, and data from only one of the surveys are used for each item in the publication tables. Therefore, when expenditure information is collected about a particular item category in both surveys, a decision needs to be made about which source of information is the more reliable for publication purposes.

In the Interview Survey, consumer units[1] (CUs) are visited once every 3 months over a period of 13 months. The survey collects expenditures on items that respondents can reasonably recall for a period of 3 months or longer, such as furniture or vehicle purchases, and expenses that occur on a regular basis, such as rent, utility payments, and insurance premiums. In the Diary Survey, CUs report expenditures in two consecutive 1-week diaries. Although CUs are asked to report all of their expenditures in the diaries, the focus of the diaries is on the expenditures of frequently purchased items.

Background

Prior to 1980, the CE was conducted at about 10-year intervals. However, since 1980 it has been conducted as an ongoing survey. From 1980 to 1983, CE data were published separately for the Diary and Interview surveys, but beginning in 1984, selected data were chosen from each survey and combined to produce publication tables.

Such integrated data from the BLS Diary and Interview Surveys provide a complete accounting of consumer expenditures and income that neither survey alone is designed to do. For example, the Diary Survey does not collect data on expenditures for overnight travel or information on reimbursements, whereas the Interview Survey does. Examples of expenditures for which reimbursements are not collected in the Diary Survey are medical care; automobile repair; and construction, repairs, alterations, and maintenance of property. The Interview Survey does not collect detailed food expenditures, or expenditures for housekeeping supplies, personal care products, and nonprescription drugs. These are items collected uniquely in the Diary Survey.

For items that are unique to one survey or the other, the choice of which survey to use as the source of data is obvious. This article briefly describes the methods employed to select the appropriate survey source for published survey estimates where there is overlap in coverage between the surveys.

[1] A consumer unit is defined as members of a household related by blood, marriage, adoption, or other legal arrangement; a single person living alone or sharing a household with others but who is financially independent; or two or more persons living together who share responsibility for at least 2 out of 3 major types of expenses—food, housing, and other expenses. The terms consumer unit and household are often used interchangeably.

Brett J. Creech is an economist in the Branch of Information and Analysis, Division of Consumer Expenditure Survey, Bureau of Labor Statistics.
Barry P. Steinberg is a Mathematical Statistician in the Statistical Methods Division, Consumer Expenditures Survey, Bureau of Labor Statistics.

Past methods of source selection

Expenditure items in the current CE are identified using the Universal Classification Codes (UCC) system. A UCC is a six-digit code that classifies reported expenditures at the most detailed level. An example of a six-digit UCC is "Tolls or Electronic Toll Passes," which is classified as UCC 520541. Selection of survey source for UCCs common to both the Diary and Interview Surveys was first conducted for tabulations of 1984–86 data through an Estimated Mean Square Error (MSE) method that used 1982–84 data. This method added the variance of the CE data to the squared difference between the mean of the CE data and the Personal Consumption Expenditure (PCE) produced by the Bureau of Economic Analysis to produce an estimate of the MSE for each CE source. The source with the smaller MSE was chosen. The method of source selection was changed in 1997 and used CE data from 1993–95. A Coefficient of Variation (CV) was computed for each source, and the source with the smaller CV was selected.

CE data are used extensively by the Consumer Price Index (CPI). In 2001, meetings were conducted at the request of the CPI group to look at differences in source selection between the CE and the CPI using 1999 data. At that time fewer than 15 UCCs had different sources between the two programs. It was recommended that the CPI adopt the CE source decision in all cases with greater than 50 reports of expenditures at the UCC level. Subsequently when new expenditure items and UCCs were introduced in 2005, source selection was coordinated so that the CE and the CPI were in agreement on the newly introduced UCCs.

In 2007, a team was formed with the task of reviewing the previous methods of source selection and developing a new quantitative methodology for selecting expenditure data from the two surveys.

CE coverage versus CPI coverage

The CPI is a measure of the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services. The CE provides BLS with expenditure data that are used to calculate the relative importance of items in the market basket. One reason the CE and CPI used different sources in the past is that the population coverage of the CE differs from that of the CPI. The CPI is calculated for urban CUs, whereas the CE uses all CUs (urban and rural) in their calculations. Definitions of components also differ between the CE and the CPI. For example, homeownership is treated differently in the two surveys: actual expenditures of homeownership are reported in the CE, whereas the CPI uses a rental-equivalence approach that estimates the change in the cost of obtaining, in the rental marketplace, services equivalent to those provided by owner-occupied homes.

The CE publishes expenditures for items such as medical care and auto repair net of reimbursements by health insurance and vehicle insurance, respectively. Reimbursements for these expenditures are captured in the Interview Survey and are used in calculating out-of-pocket expenditures. The Diary Survey does not collect reimbursement data so expenditures for these items are necessarily taken from the Interview Survey. There are also transportation UCCs that are derived from information exclusively in the Interview Survey. For example, for a new car purchase, the value of any trade-in vehicles is deducted from the purchase price to calculate the net out-of-pocket expense for the new car.

Other issues affecting source selection

A small number of UCCs are excluded from the source selection process. In some cases, the number of Diary Survey observations reported directly by respondents was so small that it disqualified the Diary Survey as a source. While the source selection methodology generally evaluates sources based on the number of expenditure reports over a given year, there are some items included in the chained C-CPI-U price index for which a sufficient number of monthly expenditure reports are required. In order to compare the monthly expenditure counts from both surveys, the Diary Survey's monthly counts have to be adjusted upward to account for the Interview Survey's longer recall period and larger sample size. During the 2007 source selection process, the source for four UCCs was based on a monthly comparison of adjusted Diary Survey counts to Interview Survey counts.

Source selection methodology

Implementing the source selection methodology involves a number of steps. Counts, sample means, and sample variances are calculated for each UCC. Before calculating means and variances, expenditure data are top coded and bottom coded to minimize the impact of outliers. Bottom coding is a form of censoring the data and is performed by applying the value of the 1st percentile to replace all smaller values. Conversely, top coding applies the value at the 99th percentile to replace all larger values for each UCC.

Next, the counts (each represents a reported expenditure for that UCC) and Z-Scores (defined below) are weighted for the three most recent collection years using the following scheme, which places greater emphasis on the more recent collection years:

- 1st collection year (oldest) by 1/6. (For 2007, 2004 data are used.)
- 2nd collection year (middle) by 2/6. (For 2007, 2005 data are used.)
- 3rd collection year (most recent) by 3/6. (For 2007, 2006 data are used.)

If a new UCC was created within the most recent 2 years or if there was a change in the collection instrument that caused a significant difference between the means in the years before and after the instrument change, then the 2 most recent years of data are analyzed. Counts and Z-Scores are

weighted with more emphasis given to the most recent collection year:

- 1st collection year (oldest) by 2/5. (For 2007, 2005 data are used.)
- 2nd collection year (most recent) by 3/5. (For 2007, 2006 data are used.)

Source selection decision criteria

Definitions of the statistical terms used in the analysis are as follows:

1) **UCC Mean**—the weighted annual average expenditure for the CPI-U population using the adjusted full sample weight.

2) **UCC Z-Score** =

$$\frac{(\overline{X}_I - \overline{X}_D) - (\mu_I - \mu_D)}{\sqrt{\sigma^2_{\overline{X}_I} + \sigma^2_{\overline{X}_D}}}$$

$\overline{X}_I$ = Annual UCC mean[2] from the Interview Survey

$\overline{X}_D$ = Annual UCC mean from the Diary Survey

$_I$ = Annual UCC population mean from the Interview Survey

$_D$ = Annual UCC population mean from the Diary Survey

$\sigma^2_{\overline{X}_I}$ = Annual UCC variance from the Interview Survey

$\sigma^2_{\overline{X}_D}$ = Annual UCC variance from the Diary Survey

With the null hypothesis that the team tested, H_o: $_I = _D$ or $\mu_I - \mu_D = 0$, the Z-Score represents the test of equality between the two weighted source means. The numerator is the difference between the sample means $(\overline{X}_I - \overline{X}_D) - 0.$ The denominator is the standard deviation of that difference; it is assumed that the two surveys are statistically independent of each other. Variances are estimated using the method of Balanced Repeated Replications (BRR)[3] with 44 replicates.

In order to determine which source to select for each UCC, the following decision criteria are used:

Criterion 1: Counts Sufficiency. For each UCC and each survey, the number of CUs with at least one expenditure is counted for each of the 3 most recent data collection years. This yields six counts for each UCC: three yearly counts for the Interview Survey and three yearly counts for the Diary Survey. These counts are used to ensure that a sufficient amount of data is available to make source selection decisions. A sufficient amount of data exists when the count for each of the 3 years is greater than or equal to 60.[4]

- If both surveys have sufficient data, then proceed to Criterion 2.
- If both surveys lack sufficient data, then keep the original source.
- If one survey has sufficient data, but the other has insufficient data, then a weighted average of the three yearly counts for the survey having an insufficient amount of data is computed: $n^* = (3/6)n_{t-1} + (2/6)n_{t-2} + (1/6)n_{t-3.}$[5]
- If the weighted average n^* from the insufficient survey is greater than or equal to 60, then proceed to Criterion 2.
- If the weighted average n^* from the insufficient survey is still less than 60, then use the survey with sufficient data as the source.

Criterion 2: Statistical Significance.

a) If the absolute value of the weighted Z-Score, $z^* = (3/6)z_{t-1} + (2/6)z_{t-2} + (1/6)z_{t-3}$, is greater than or equal to 1.645 then select the source based on the following[6]:

- If the weighted Z-Score is greater than or equal to 1.645, then the Interview Survey is selected as the source.
- If the weighted Z-Score is less than or equal to –1.645, then the Diary Survey is selected as the source.

b) If the weighted Z-Score is between –1.000 and 1.000, then the current source will continue to be used.

c) If the weighted Z-Score is greater than 1.000 and less than 1.645 or less than –1.000 and greater than

[2] $\overline{X}_I$ and $\overline{X}_D$ represent the weighted means of collected data from the Interview and Diary Surveys, respectively. $_I$ and $_D$ represent the population means for the Interview and Diary surveys, respectively, which are unknown. For the Z-score calculation, the null hypothesis tests that the difference between $_I$ and $_D$ is zero.

[3] Balanced Repeated Replication (BRR) is a method of variance estimation used for sample survey statistics when the complexity of a survey's sample design prevents standard classical variance estimation techniques from being used. BRR belongs to a class of variance estimation techniques that use *replications*. The basic idea behind replication is to select subsamples of the collected data repeatedly from the full sample, and then calculate the statistic of interest from both the full sample and from each sub-sample. These sub-samples are called *replicates*. The difference between the replicate estimates and the full sample estimate are then used to estimate the variance of the full sample statistic.

[4] The number 60 represents an average of five expenditures per month, which was found to be the minimum number of expenditures needed to produce stable results.

[5] For the equation $n^* = (3/6)n_{t-1} + (2/6)n_{t-2} + (1/6)n_{t-3,}$ t represents the evaluation year with the collection years being t–1, t–2, and t–3. For example, a 2007 evaluation will have t–1 representing collection year 2006, t–2 representing collection year 2005, and t–3 representing collection year 2004. The same concept applies to $z^* = (3/6)z_{t-1} + (2/6)z_{t-2} + (1/6)z_{t-3}$ that will be mentioned later in the article.

[6] The value of 1.645 represents the 95th percentile of the standard normal distribution. It is often used in research as the critical value in determining statistical significance. On the left side of the standard normal distribution, the value of –1.645 represents the 5th percentile and is used in a similar manner.

–1.645, the following method is used to select the source:

- If all three Z-Scores are greater than 1.000, then the Interview Survey is selected as the source.
- If all three Z-Scores are less than –1.000, then the Diary Survey is selected as the source.
- In all other scenarios, the source remains the same.

2007 data results from the source selection process

Table 1 lists the number of overlapping UCCs researched for source selection and the number of UCCs that changed sources. Out of approximately 900 UCCs, there were 222 overlap UCCs that were researched, resulting in 22 changing sources. Eighteen changed because they had high Z-Scores (absolute values) and 4 changed due to the Diary Survey failing the counts criterion, thereby switching to the Interview Survey. There were a total of 9 UCCs that had fewer than 60 observations in both surveys; therefore the source stayed the same. A total of 75 UCCs had observations in the Diary Survey that failed the counts criterion, thereby using the Interview Survey as the source. Only one UCC had an observation in the Interview Survey that failed the counts criterion. There were a total of 94 UCCs that had high Z-Scores (their absolute values were greater than or equal to 1.645), 27 UCCs with low Z-Scores (their absolute values were less than 1.000) and 16 UCCs with Z-Scores between 1.000 and 1.645 or between -1.000 and –1.645.

2009 results

The source selection process implemented in 2007 was used to evaluate 2006–08 data for the development of 2009 estimates. Upon receiving a list of new UCCs, the source selection program was run on all UCCs having data in both the Interview and Diary surveys. The 3-year period from 2006–08 was used for most UCCs, and 2007–08 data for the newer UCCs. After calculating the expenditure counts and Z-Scores, the procedures identified six UCCs for which the applicable source selection criterion had changed. After further evaluation the source was changed for only two of them.

The source for two of these six UCCs had been changed in 2007 and it was decided not to change again in 2009, to avoid an undue switching between sources every 2 years. Of the remaining four UCCs, it was decided to keep the Interview Survey as the source for two UCCs because of concerns about the consistency of the Diary Survey estimates. When the analysis was completed, two UCCs changed sources from the Diary Survey to the Interview Survey.

Summary

Source selection is the process of choosing the better survey to use in CE's official published expenditure estimates. It is a multistep process performed every 2 years for every overlapping UCC by comparing expenditure data from both the Interview and Diary Surveys using a counts criterion and weighted Z-Score approach to determine the better source for use in CE production. The method uses the previous 3 years of data when available, giving more weight to the most recent years. For new UCCs only 2 years of data are used. The data are adjusted for outliers in both the Interview and Diary Surveys. A number of criteria are then tested to determine which source to select. The first criterion assesses the number of unweighted consumer units making an expenditure for each UCC in each survey and may eliminate a source where an insufficient number of CUs report. The next criterion chooses the source that provides the larger overall expenditure per UCC. The means of reported expenditures, weighted by year, are compared from each survey using a standard Z-test, and in essence, the statistically significant larger mean is chosen.

The source selection process will continue to be run every 2 years. In addition to running the process, the CE will be looking for ways to accelerate the process by methods such as automation. Future research will also be performed to adapt the process for changes in survey instruments, collection methodology, and data processing. ■

Table 1. **Number of overlap Universal Classification Codes (UCC) researched for source selection**

	UCCs researched	UCCs changing source
Total	222	22
UCCs for any year with observations less than 60 in both surveys	9	0
UCCs in which the observations in the Diary survey fail the counts criterion	75	4
UCCs in which the observations in the Interview survey fail the counts criterion	1	0
UCCs with Z-scores between –1.00 and 1.00	27	0
UCCs with Z- scores of +–1.00 and +–1.645	16	0
UCCs with Z- scores of –1.645 or less	68	17
UCCs with Z- scores of 1.645 or greater	26	1