# Components of Error Analysis in the Current Employment Statistics Survey[1]

Larry Huff and Julie Gershunskaya
U.S. Bureau of Labor Statistics, 2 Massachusetts Ave NE, Suite 4985, Washington, DC, 20212

**Abstract**
One of the prominent features of the Current Employment Statistics (CES) survey conducted by the U.S. Bureau of Labor Statistics (BLS) is that the data collection is conducted monthly and the estimates are published in a very short time period. It is important to analyze the effect of the survey nonresponse on the quality of the estimates. Several months after the reference period, universe data become available through the Quarterly Census of Employment and Wages (QCEW) BLS Program. The availability of the census data makes it possible to separate sources of potential divergence of the CES estimates from the corresponding population totals. The sources include sampling variability; the potential for nonresponse bias; the differences in the data reported to the two programs, i.e. reporting error; and frame imperfections due to the appearance of new establishments (births).

**Key Words:** nonresponse bias, error decomposition, response rate

## 1. Introduction

Current Employment Statistics (CES) is a large-scale establishment survey conducted by the U.S. Bureau of Labor Statistics. The CES estimates of employment and other important indicators of the U.S. economy are published monthly at various levels of industrial and geographic detail.

The CES estimates are affected by a composition of errors due to sampling variability, potential nonresponse bias, measurement error, and frame imperfections. The purpose of this study is to evaluate the degree to which the different components of error affect the resulting estimate of the employment level.

We base the research on historical universe administrative data from the Quarterly Census of Employment and Wages (QCEW) BLS program. This data, along with the CES reported data, can be used to decompose the total error into its parts.

For this paper, in particular, we focus on the evaluation of the nonresponse bias and its possible relationship to the unit response rate.

In what follows, we give a brief overview of the CES survey and the relevant details of the QCEW data. Section 2 contains the description of the CES error decomposition study based on the QCEW data. The outcome of the study is presented in Section 3. In Section 4 we discuss the results and entertain an alternative approach to the bias assessment based solely on the CES data. The last Section contains the summary.

### 1.1 Sample design

The CES uses a stratified simple random sample of about 200,000 unemployment insurance (UI) accounts, consisting of over 800,000 establishments. A frame for the CES sample selection is created from the QCEW data file. Strata on the frame are defined by State, industrial supersector

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

based on the North American Industry Classification System (NAICS) and on the total employment size of establishments within a UI account.

## 1.2 QCEW data

The QCEW file is based on administrative Unemployment Insurance tax data containing a near census of U.S. businesses. After a lag of approximately 6 to 9 months, the QCEW essentially provides the employment counts that the CES survey "predicts" in real time. The QCEW data plays an important role in the CES survey: it provides the sampling frame, the benchmark levels, and a valuable historical source of data used for research purposes. It has, however, some important drawbacks, besides being not as timely as the sample based data, because of its administrative nature and the quarterly reporting pattern.

## 1.3 CES Data collection features

Timeliness of the estimates is a special feature of the CES survey. The data collection is performed on a very tight schedule and about 260,000 CES reports are collected each month.

For the 1st (preliminary) closing especially, the data collection is challenging: CES has only 10 to 16 days to collect the data (the number of collection days varies because the *Employment Situation* release date is scheduled around the Current Population Survey production schedule). After the 1st closing too, there are only about two weeks of follow-up for non-respondents before collection for the next month must begin.

Data collection is accomplished using several modes, including, Electronic Data Interchange (EDI), Computer Assisted Telephone Interview (CATI), World Wide Web, FAX, Touchtone Data Entry (TDE), Mail, and Special Arrangement[1]. Data collection centers (DCC) have a comprehensive program for enrollment of new sample units and refusal conversion.

To facilitate the response, units may choose to provide data as an aggregate figure for a number of worksites or they may report for individual establishments.

See also Rosen *et al.* (1999).

## 1.4 The form of the estimator

Each month, CES estimates the relative growth of employment from the previous to the current month. The estimates are derived for each elementary estimation cell. The numerator of the ratio is the survey weighted sum of the current month reported employment; similarly, the denominator is the survey weighted sum of the previous month employment. Only establishments reporting positive employment in both months are used in the estimation (a matched sample of establishments):

$$\hat{R}_{c,t} = \frac{\sum_{j \in M_{c,t}} w_j y_{j,t}}{\sum_{j \in M_{c,t}} w_j y_{j,t-1}}, \qquad (1.1)$$

[1] Special Arrangement of the data collection is an example of the BLS effort to increase response. BLS is willing to accept information in any manner that businesses are willing to report it, even if they have refused to report in one of the many ways set up to facilitate reporting. It means that the units may respond by sending reports in a non-standard format, for example, as a computer printout. BLS is then responsible for keying the data into a format it needs.

where $w_j$ is sampling weight of establishment *j*; $y_{j,t}, y_{j,t-1}$ employment reported by establishment *j* at months *t* and *t-1*, respectively; $M_{c,t}$ denote a set of establishments in cell *c* that report positive employment in both months *t* and *t-1*.

Thus, $\hat{R}_{c,t}$ is the estimate of the relative employment change in the continuous part of the population of establishments. CES cannot directly account for appearance of new businesses ("births") and uses a model to compensate for this deficiency. Research has shown that the employment "lost" due to the out-of-business units ("deaths") is a reasonable approximation of the amount of employment gained from births units (Butani 1997). This is the underlying assumption for using (1.1). Thus, by "imputing" for deaths, we estimate the employment due to births. The above assumption, however, does not hold exactly, and a relatively small correction, the net difference between the births employment and the imputed (deaths) employment, is modeled using the historical data (see Kropf *et al*. 2002). Every month, the additive adjustment ("Net Births-Deaths") is used in the formula for the employment levels, as follows:

$$\hat{Y}_{c,t} = \hat{Y}_{c,t-1}\hat{R}_{c,t} + \hat{N}_{c,t}, \qquad (1.2)$$

where

$\hat{Y}_{c,t}, \hat{Y}_{c,t-1}$ are the employment levels in the estimation cell *c* at months *t* and *t-1*, respectively; $\hat{N}_{c,t}$ is the model-based Net Births-Deaths (Net BD) adjustment for cell *c* at month *t*.

The estimate of employment at a higher level of aggregation (denoted as *C*) is simply the sum of the basic estimation cell levels:

$$\hat{Y}_{C,t} = \sum_{c \in C} \hat{Y}_{c,t} \,. \qquad (1.3)$$

This chain of level estimates starts at a point in the past – the starting point (March of each year) is a census level figure derived from the QCEW. Once a year, when the March QCEW figure becomes available, BLS also revises the estimates derived previously from the CES sample to conform to the known census level. See the *BLS Handbook of Methods* (2004, Chapter 2) for further details.

## 1.5 Components of error in the estimates

The CES estimates are affected by a composition of errors, each having its specific impact on the result. First, it is *sampling error*, reflecting the fact that the estimates are based only on part of the universe rather than the whole census. Because we are using a probability sample, the sampling error contributes only to the variance of the estimator.

Second, is the error due to the fact that not all units respond to the survey. This type of error affects the variance; but it also may contribute to the bias of the estimate in cases where the responding units are different from non-respondents. We call this component *the nonresponse error*.

There are differences in the data reported to CES and data reported by the same set of units as recorded in the QCEW file. If, despite the shortcomings of the QCEW data, we regard it as "truth", then we can tentatively call the third type of error *the reporting error*. Some of the CES-QCEW differences are related to the nature of the QCEW data that are quarterly administrative data with all possible consequences. The differences also can be explained in part by the fact that some units report to CES in an aggregate manner, i.e., for multiple establishments within a UI account.

As mentioned earlier, CES cannot directly account for business births. The fourth source of error is due to the imperfections of the sampling frame, which is inevitably out of date – in that the sampling frame cannot contain new (birth) businesses on a concurrent basis – at the time the sample is implemented for estimation. To compensate for this *frame inadequacy*, i.e., missing business births, CES uses a model (as described in subsection 1.4). The model error potentially contributes to the estimate bias and variance.

## 2. Design of the study

The approach to the error decomposition used in this paper is similar to the one described in Gershunskaya *et al*. (2002). We now briefly review the approach.

The QCEW data file can be used to extract data for the complete set of sampled units. Their data recorded in the file for a required period of time (here, one year of data) can be used to simulate the estimates. The purpose of this simulation is to show what would be the sample based estimates if the whole sample were available (i.e., without nonresponse) to the survey. We call the series of these estimates over the given period of time *the whole-sample line*.

Each month, the estimate of the relative trend targets the part of the population that consists of all QCEW units reporting positive employment in the previous and current months. Thus, *the target continuous population line* can be derived similar to the sample line as if we observed the whole population and each population unit had the weight of one.

Instead of using all sampled units, we can derive the estimates based on the subset corresponding to only those units reporting to CES. The estimates are then simulated using the QCEW data and based on this subset of the CES-respondents. This series of estimates is called *the respondents line*. The difference between the whole-sample line and the respondents line shows how the estimates are affected by the nonreponse. Any pattern found in the values of these differences would be useful in revealing the nonresponse bias.

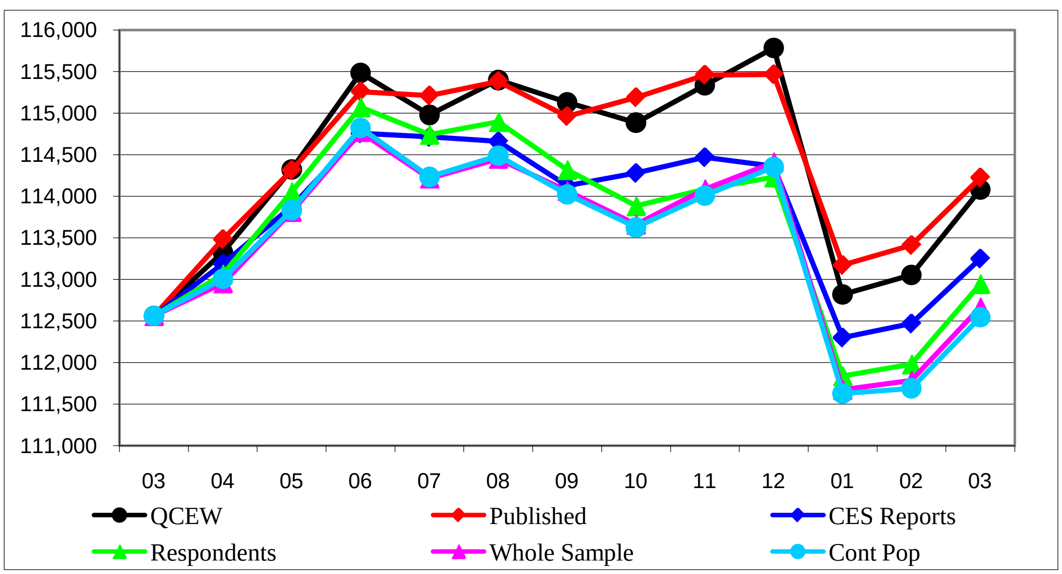


**Figure 1:** Estimates at the National Total Private level for March 2006 – March 2007 period.

Let us call the series using the CES data, *the CES-reporting line*. The difference between the CES-reporting line and the respondents line using QCEW data is due to the difference in the data reported to the CES and QCEW by the same set of units, we call this difference the reporting error.

The plot of the estimates for the period from March 2006 (the benchmark point) through March 2007 is presented in Figure 1 to illustrate the description. The estimates on the plot were calculated using National supersector level as the basic estimation cell.

First, note that all lines start from the same point in March 2006. This is a benchmark level, a starting point, used in the CES annual cycle of estimates.

The black line displays the true population levels as derived from the QCEW data[1]. Close to the black line, is the red line displaying the published CES estimates. The four lines that are significantly lower than these two lines correspond to the whole-sample, the respondents, the CES-reports, and the target continuous population lines. The reason these lines are so much lower than the published line is that they do not include the Net BD adjustment.

## 3. Results

Since the QCEW and CES data have different seasonal patterns, we focus only on the comparison of the final, 12-month after the benchmark, estimates. We present the results from the several years of study and the analysis of the relation of the nonresponse (NR) bias and the response rates (RR).

The differences between the estimates 12 months after the benchmark are shown in Table 1[2].

| | ***2004*** | ***2005*** | ***2006*** | ***2007*** | ***2008*** |
|---|---|---|---|---|---|
| **NR** | 218,600 | -206,200 | 28,900 | 280,400 | 273,300 |
| **SE** | -198,000 | 447,400 | 218,600 | 120,100 | 29,800 |
| **RE** | 136,500 | 153,100 | -105,400 | 305,000 | 578,200 |
| **B/D[3]** | -526,900 | -350,100 | -766,100 | -482,800 | -465,700 |
| **Total** | -369,800 | 44,200 | -624,000 | 121,500 | 415,600 |

**Table 1:** Summary of Errors, National Total Private level: NR = (respondents) – (whole-sample); SE = (whole-sample) – (target continuous population); RE = (CES reports) – (respondents); B/D = (Net BD added) + (target continuous population) – (QCEW level).

The total NR error does not appear to be consistent in direction or particularly related to the response rates. (The response rates have generally but marginally gone down over this 4 year period).

[1] Included in the levels are corrections for the units not covered by QCEW, i.e., not-elsewhere-classified (NEC) and presumed non-covered (PNC).

[2] The table is based on the research version of the CES estimates (computed at the National supersector level as the basic estimation cell) rather than on the officially published CES estimates. As a result, the "Total" line in Table 1 is an approximation suitable for the purposes of the current research. It should not be confused with the officially reported benchmark revisions of the CES estimates.

[3]Note that this line does not reflect an error in the Net B/D model adjustment used in production. This line indicates the difference between the cumulative adjustment that was actually added to CES estimates during the 12 month cycle and what would have to be added to the *QCEW based continuous population* level to bring it up to the QCEW level. However, because of the observed persistent differences in the CES and QCEW patterns and levels, the amount that "truly" has to be added to the estimates is difficult to assess based on this type of analysis.

The NR error is not a principal driver of the total survey error: the NR error is not substantially larger than the other components of error for CES and the year where the NR error is smallest (2006) had the largest total error.

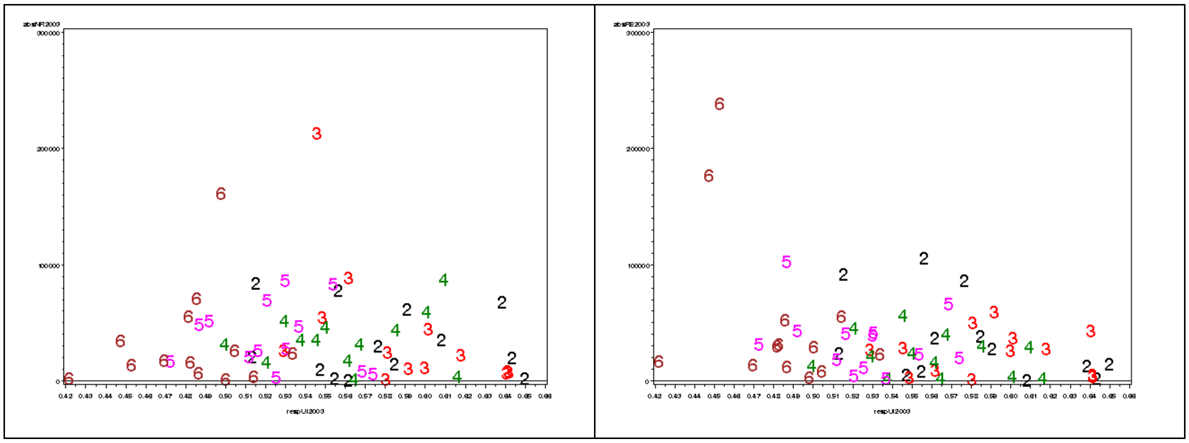

**Figure 2:** (a) Absolute NR bias vs. Response Rate. (Estimates are shown at the supersector level. Different numbers refer to different years, e.g., "6" refers to 2006 sample frame/2007 benchmark year); (b) Absolute Reporting Error (RE) vs. Response Rate. (Estimates are shown at the supersector level. Different numbers refer to different years).

Figure 2a shows the absolute value of the Nonresponse Error versus the Response Rate. Each data point corresponds to a supersector and the numbers signify the year of the sampling frame (from "2", for the 2002 frame, to "6" for the 2006 frame). We can observe that the response rates are lower in recent years (4, 5, 6) than in the earlier years. However, there is no evidence that the lower response rate is associated with the higher nonresponse bias.

On the other hand, the Reporting Error (RE) is large for two industries (Business Services and Leisure) that have low response rate (Fig.2b). These two outlying points may be an indication of a tendency of the higher reporting error to be associated with the lower response rate. This needs to be monitored further.

How can we explain this effect on the reporting error? As noted earlier, the nature of the differences between the CES and QCEW reports is such that the naming of the sources of error is only tentative. For example, one notable difference between QCEW and CES is that the CES response may include the expansion establishments (i.e., the new establishments that are part of an existing sampled UI account). If the movement of the expansion units is different from the movement of the "older" establishments and if it is more likely that respondents have more expansion units than nonrespondents, then the bias occurs. However, this potential difference between respondents and nonrespondents cannot be detected by examining the QCEW-only-based nonresponse error (e.g., Fig 2a), because the QCEW-based estimates, by definition, do not include any expansion units. Because it is not realistic to assume that nonrespondents do not have the expansion units at all, we do not know, based on the QCEW-based analysis if the CES respondents indeed differ from nonrespondents. The bottom line is that, although related, the CES and QCEW data come from two different stochastic universes. Thus, the results of error decomposition relying on the QCEW data need to be interpreted cautiously.

In the next Section, we use only the CES data to explore the effect of the aggregation structure on the CES estimates.

## 4. Effect of the aggregation structure on CES estimates

The implicit assumption used in the CES estimation formula (1.1) is that the units within a given basic estimation cell have equal probabilities to respond. The basic cell structure is different for the State and National level estimates. The State level estimation cells are based on a higher level industry while the National cells are defined at very detailed industrial levels without regard to geography.

The risk is that the higher level industry cells are not as homogeneous in terms of the growth rate. If subindustries with differential relative growth also differ in their units' probability to respond, the estimate derived at the aggregate level industry may diverge from the estimate obtained by adding up from the lower level cells. Similarly, the problem may arise if States have differential growth rates in a given subindustry and also have different rates of response.

Thus, if there is enough sample, making estimates at detailed industrial levels and aggregating them up to a higher level would be preferable because the lower level cells are usually more homogeneous than the combined cells. At the State level, however, the sample often is not large enough to make reliable sample based estimates at levels lower than the supersector level. If the State level estimates are produced using supersectors as the basic estimation cells while the National estimates are made at the National detailed industry levels and then aggregated up, then the discrepancy between the National estimates and the sum of States estimates is possible and it would indicate nonresponse bias in either or both of the estimates.

To explore a possible effect of the nonresponse on the differences between the sum of States estimates and the sum of the National cell estimates, we computed the estimates of trends (1.1) using three versions of the basic cell definition. Cells were defined using (i) 4-digit level of the NAICS code (Na4d); (ii) industrial supersector level (NaSS); (iii) intersection of a State and industrial supersector (StSS). Then we obtained the National Total Private estimate (less the Net BD) using (1.2) and (1.3) (for the purposes of this study, we ignored the Net BD adjustment, i.e., set $\hat{N}_{c,t} = 0$ in(1.2)).

If there is no bias due to nonresponse, the three versions of the Total Private estimate should be approximately the same. If State is an important factor in the nonresponse bias adjustment, then estimates aggregated from the StSS levels would diverge from the estimates obtained from the NaSS cells. If the detailed industry plays the role, then Na4d version would be different from NaSS.

We obtained the following results, for the five years that we studied. The cumulative differences between the series, from April through March of the following year, for the five years of study, are shown in Table 2.

In some months, the differences between Supersector based series and the 4-digit based series may accumulate to a substantial amount (e.g., -168,062 from March to October 2004 and 148,990 from March to July 2007). By contrast, differences between the State-based and National-based series range only from -54,685 to 46,480, which can be tolerated for estimates at this high level. In a few larger industries, this effect is more pronounced; however, it is moderate to negligible in other, smaller industries.

Thus, the discrepancies mainly result from the differences in the basic cell structure due to the industrial detail (supersector vs. 4-digit NAICS) rather than geography (State supersector vs. National supersector) suggesting that the State dimension in the definition of the National cell structure is not as important as the industrial detail.

| | *2003-2004* | | *2004-2005* | | *2005-2006* | | *2006-2007* | | *2007-2008* | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **StSS-Na4d** | **StSS-NaSS** | **StSS-Na4d** | **StSS-NaSS** | **StSS-Na4d** | **StSS-NaSS** | **StSS-Na4d** | **StSS-NaSS** | **StSS-Na4d** | **StSS-NaSS** |
| **Apr** | 6,900 | -3,408 | -2,154 | 6,442 | 62,092 | 10,921 | 6,708 | 5,490 | 54,589 | 15,739 |
| **May** | -1,123 | -14,439 | -37,132 | -3,858 | 40,135 | 5,039 | -14,157 | 1,146 | 73,939 | 16,994 |
| **Jun** | 9,194 | 7,325 | -31,756 | 722 | 26,174 | 5,761 | -30,014 | 8,949 | 131,154 | 46,480 |
| **Jul** | 40,125 | 11,372 | -44,059 | -9,567 | 36,125 | -10,297 | -37,892 | -17,688 | 148,990 | 31,012 |
| **Aug** | 35,162 | -6,911 | -79,600 | -19,382 | -5,922 | -14,728 | 1,468 | -18,817 | 134,503 | 26,431 |
| **Sep** | -88,689 | -20,498 | -128,419 | -41,382 | -95,421 | -54,685 | -65,645 | -37,064 | 60,248 | 12,114 |
| **Oct** | -126,040 | -32,834 | -168,062 | -46,669 | -107,882 | -43,472 | -89,674 | -31,556 | 36,085 | 25,749 |
| **Nov** | -96,570 | -39,561 | -108,654 | -22,511 | -110,762 | -29,514 | -85,869 | -22,407 | 25,876 | 41,857 |
| **Dec** | -114,066 | -34,960 | -99,619 | -9,838 | -110,913 | -12,555 | -72,935 | 1,697 | -7,209 | 23,462 |
| **Jan** | -83,836 | -3,936 | -75,449 | -12,112 | -58,439 | 2,558 | -42,718 | -14,323 | 21,345 | 24,169 |
| **Feb** | -98,668 | -9,062 | -133,912 | -26,035 | -52,203 | -11,850 | -72,824 | -42,793 | -14,179 | -4,442 |
| **Mar** | -127,035 | -13,854 | -113,832 | -28,310 | -39,590 | 223 | -57,634 | -33,123 | 5,850 | -3,588 |

**Table 2:** Cumulative differences between series obtained using as basic estimation cells National supersectors (NaSS), 4-digit NAICS levels (Na4d), and State supersectors (StSS).

## 5. Summary

The QCEW data is a valuable source for the posterior analysis of the sources of error in the CES estimates. However, there are differences between the CES and QCEW data and, as a consequence there are significant differences between the respective sample trends. This causes difficulties in interpreting the results of the error decomposition based on the QCEW data.

Analysis using the CES data shows that estimates depend on the aggregation structure. This is due to the fact that different basic estimation cells have differential response rates and rates of employment growth. Nonresponse bias adjustment for the State level estimates may improve the estimates and reduce the discrepancy between the Sum of States and National estimates.

## Acknowledgement

We would like to thank Dan Zhao for the programming support of this project.

## References

Bureau of Labor Statistics (2004), Chapter 2, “Employment, hours, and earnings from the Establishment survey,” *BLS Handbook of Methods*. Washington, DC: U.S. Department of Labor

Butani, S. (1997) “Analysis of Births and Deaths Employment Estimates.” *Internal BLS/OEUS/SMS Research Paper*

Gershunskaya, J., Eltinge, J.L. and Huff, L. (2002) "Use of Auxiliary Information to Evaluate a Synthetic Estimator." *Proceedings of the Survey Research Methods of the American Statistical Association*

Kropf, J., Strifas, S. and Traetow, M. (2002) "Accounting for Business Births and Deaths in CES: Bias vs. Net Birth/Death Modeling." *Proceedings of the Survey Research Methods of the American Statistical Association*

Rosen, R., Harrell, L., Manning, C., and Skuta, D. (1999) "Data Collection Issues Related to Implementing the Redesigned Current Employment Statistics Survey," *Proceedings of the Section on Survey Research Methods, American Statistical Association*