Determining the Goals and Effectiveness of the Consumer Expenditure Survey's Calibration Procedure

Brian T. Nix
U.S. Bureau of Labor Statistics, 2 Massachusetts Avenue NE, Washington, DC 20212

Abstract
"Calibration" is a weighting procedure performed by many household surveys that modifies the sample households' weights to make them add up to certain known population values. In practice the number of known population values must be small to avoid small sample size problems like empty cells. When the number of known population values that are available exceeds the number that are usable, a decision needs to be made about which ones to use. Naturally, one wants to use the best set of known population values, but that requires knowing what the word "best" means. In this paper we review what various authors consider the goals of the calibration procedure to be, and how these goals apply to the Consumer Expenditure Survey.

Key Words: Calibration, Consumer Expenditures, Survey Methodology, Weighting

1. Introduction

The goal of the Consumer Expenditure Survey is to find out how Americans spend their money. To accomplish this goal, the CE Survey[1] uses the data from its survey respondents to estimate the average annual expenditure per household for the whole country, and for various subgroups of households in the country.[2] Part of the estimation process involves calculating weights for the respondent households in a way that makes them add up to the total number of households in the country, and also to the total number of households in various subgroups.

Like most household surveys, the CE Survey has a three-step weighting process. In the first step every household in the sample is given a base weight equal to the inverse of its probability of selection. In the second step, the base weight of every respondent household is multiplied by a nonresponse adjustment factor equal to the sum of the base weights of the eligible households divided by the sum of the base weights of the respondent households.[3] And in the third step, the nonresponse-adjusted weights of respondent households are multiplied by a calibration adjustment factor equal to the known population of the country divided by the sum of the nonresponse-adjusted weights of the respondent

---

[1] The CE Survey, sometimes called the CE *Surveys*, initially consists of two separate, independently collected surveys; these are a four-panel Interview Survey and a two-week Diary Survey, the results from which are then combined prior to publication.
[2] More precisely, "consumer units" (CUs), which are members of a housing unit, often related, who share income resources and make joint expenditure decisions; however, the more familiar term "households" will be used for the remainder of this text.
[3] "Eligible" households are those that live at a sample address and are in the survey's target population. Vacant, nonresidential, and nonexistent addresses are not eligible, as are housing units that are occupied by people outside the survey's target population.

households. This calibration process makes the respondents' final weights sum to the known population of the country.[4]

The calibration procedure is the focus of this paper.

2. The Goals of Calibration

Calibration is a weighting procedure performed by many household surveys that modifies the sample households' weights to make them add up to certain known population values. Each household begins with a probability-sampling weight, which is then adjusted to accomplish several goals.

The idea of calibration goes back to Deming and Stephan (1940), who used a method called "raking ratio estimation." They were trying to analyze a cross-tabulated sub-sample of data from the 1940 census where they knew the population's true marginal totals, but they did not know the population's true totals for the individual cells in the table. Deming and Stephan estimated the true totals for the individual cells using a method of iterative proportions. They said under ideal conditions their method would reduce variances, but they never actually advocated it. They just said it was one possible method, and that other methods could probably be used. Thus, Deming and Stephan left the precise goal of the calibration procedure undefined. Moreover, a literature review quickly reveals there is still no consensus on its precise goal.

Kott (2006, 2012) and Särndal (2007) listed several possible motivations for calibration weighting:

1. Decreasing variance,
2. Reducing or eliminating bias,
3. Forcing estimates to be consistent with external data, and
4. Correcting the sampling weights to account for an imperfect sampling frame.

---

[4] The CE Survey uses the traditional cell method in its nonresponse and calibration procedure. In the nonresponse adjustment process, the CE Survey's sample is partitioned into 24 cells based on the region of the country, the number of people living in the households, the average income of the households, and the number of contact attempts. In the calibration adjustment process, the CE Survey's sample is partitioned into 35 cells based on the division of the country, the urban/rural status, the homeownership status, the Hispanic status, and the age and race of people. The nonresponse adjustment process is performed on each of those 24 cells, and the calibration adjustment process is performed on each of those 35 cells. Thus the final weights of the respondent households sum to the known population of the U.S., and also the known population of 35 subsets of the population.

The 35 known population counts from CPS used for calibration in the CE Survey are as follows:

- (1) Total U.S. households
- (1) Total U.S. homeowner households
- (14) Populations for all combinations of 7 age categories and 2 race categories
- (9) Total populations for each of the 9 Census divisions
- (9) Urban-only populations for each of the 9 Census divisions
- (1) Total U.S. Hispanic population

Previously there were 24 counts from CPS; the addition of total Hispanic population and a transition from 4 Census regions to 9 Census divisions has not yet been reflected in the BLS Handbook of Methods.

The first of these goals of calibration is decreasing variance. Decreasing variance is a standard goal in statistics, as it results in smaller confidence intervals for the calculated estimates. While decreasing variance is usually thought of happening by means of using a larger sample size or changing the sample design, variance can also be decreased by adding new information to an estimation process. If useful auxiliary information is available, this information can thus be incorporated into the model to reduce the variance. The "calibration problem" can be seen as an optimization problem which seeks to improve initial weights in a system by incorporating this auxiliary information.

The second of these goals of calibration is reducing or eliminating nonresponse bias. Reducing bias is another standard goal in statistics, as it brings the estimated statistics from the survey closer to the true statistics of the universe which the survey represents.

The third of these goals of calibration is forcing estimates to be consistent with external data. This is an inherent feature of the calibration process, where we set weights to calibrate to known population totals. The goal therefore is not to make this happen, but to carefully select which known totals should be used for calibration as this happens.

The fourth of these goals of calibration is correcting the samplings weights to account for an imperfect sampling frame. This goal can also be expressed as adjusting for coverage errors. Calibration is used to correct sampling frame problems such as undercoverage (Dever, 2008). Undercoverage occurs when the sampling frame fails to account for all subgroups of the target population. Less frequently occurring, overcoverage, when a subgroup is overrepresented in a sampling frame, can also be corrected by calibration. Kott (2006) said that the totals used in calibration must be free of error (or to have very little mean squared error relative to the totals) in order to be used for adjusting for coverage errors.

2.1 Calibration for Decreasing Variance in the CE Survey

The first of these goals of calibration is decreasing variance. It is a standard goal among statisticians because it reduces the length of the confidence intervals of the survey's estimates. There are at least three ways of decreasing the variance of a survey's estimates. One way is to increase the survey's sample size. A second way is to change the survey's sample design. And a third way is to increase the amount of information that is used to produce the survey's estimates. Calibration uses the third way of reducing the variance of a survey' estimates. It can be thought of as a Bayesian idea, where more information leads to smaller variances:

$$V(\bar{x}_n|\theta) \leq V(\bar{x}_n)$$

As previously mentioned, the CE Survey calibrates its weights to 35 known population counts from the Current Population Survey. In mathematics it is well known that a system of equations with more variables than equations will have an infinite number of possible solutions. As a result, there are an infinite number of possible sets of calibration factors that make the weights add up to these 35 population counts, because each respondent household gets a calibration factor and there are many more than 35 respondent households.

The CE Survey takes the approach of choosing such factors that minimize the amount of change made to the nonresponse-adjusted weights – that is, the weights already determined by the nonresponse adjustment process. In this way, the goals of calibration are met with minimal change in the weights. This ideally ensures that the weights within the geographic and nonresponse cells – identical to each other prior to calibration – will remain similar after calibration, minimizing variance as much as possible.

The CE Survey uses a modified approach to the Generalized Least Squares (GLS) function. Specifically, the CE Survey finds the calibration weights w_i which minimize the weighted sum of squares

$$\sum_{i \in s} a_i \left(\frac{w_i}{a_i} - 1 \right)^2$$

where values of a_i are the nonresponse-adjusted weights, values of w_i are the calibration-adjusted weights, and the sum is over all respondent households in the sample. This yields calibration weights w_i that are as close to their nonresponse-adjusted weights a_i as possible in relative terms.

Zieschang (1990) found that applying the GLS weighting approach to the CE Survey was helpful in reducing the variances for nearly all item categories in both components of the CE Survey – the Interview and Diary surveys. For example, in the Interview Survey it reduced the CVs in the all-items category by nearly 30 percent. Thus, Zieschang found the GLS weighting approach to be successful in reducing the CE Survey's variances.

However, another way of looking at the matter is that the GLS weighting approach may not have reduced the survey's variances at all. It may have reduced the estimates of the variances, but it may not have reduced the true variances themselves. It can be argued that once a survey has a sample design and a sample size, and once it has collected its data, the survey's variances are fixed number whose values cannot be changed. They can be estimated, but they cannot be changed. Under this way of looking at things, the GLS weighting approach may have actually under-estimated the CE Survey's variances instead of reducing them. Therefore, the GLS weighting approach may not have been as successful in reducing variances as Zieschang thought.

2.2 Calibration for Reducing Bias in the CE Survey

In processing the CE Survey, the goal of reducing bias is expected to already be accomplished by the nonresponse adjustment process prior to the calibration process,[5] although there are indications that this is not fully successful. Traditionally, variance and bias are seen as a trade-off – that is, reducing one necessarily increases the other. However, this is not always true. There is nothing in statistical theory that causes an increase in one to always result in a decrease in the other. Variance and bias can move in the same direction or they can move in different directions. Lundström and Särndal (1999) pointed this out when they explained that through a prudent choice of auxiliary information in the calibration process, it is possible for both the variance and the nonresponse bias to decrease together.

We have already seen that the calibration approach furthers the work of the nonresponse bias adjustment in order to better approximate the total number of households, while also explicitly forcing the weights to be consistent with the other total population counts from the Current Population Survey.

[5] Currently, the nonresponse (or noninterview) adjustment process for CE includes four variables: Census Region, IRS Income Group, CU Size, and Number of Contact Attempts.

However, over the past several years, statisticians at the Bureau of Labor Statistics examined the CE Survey to determine the level of nonresponse bias present in the survey. The conclusion was that the nonresponse data were missing at random (MAR), and there is no significant nonresponse bias present in the survey (Steinberg *et al*, forthcoming).

As stated by Deville and Särndal (1992), we want the calibration adjustments to modify the previous weights as little as possible. This is because we presume that the initial base weights, as well as the nonresponse adjustment weights, give us unbiased population and expenditure estimates. Any significant change from those weights would therefore generate bias.

Furthermore, there are an infinite number of weights which satisfy this equation, called the *calibration constraint*

$$\sum_{i \in s} w_i \mathbf{x}_i = \mathbf{x}.$$

where the sum of the products of all values[6] $\mathbf{x}_i$, when multiplied by the weights w_i, must equal a given total[7] **x.**

Therefore, we can choose the weights which are the closest to the previous weights and still satisfy the calibration constraint. This will therefore minimize any increase in variance to the system, while still satisfying the additional constraints of the known population counts.

2.3 Calibration for Forcing Estimates to be Consistent with External Data in the CE Survey

As will be shown in this section, the nonresponse adjustment process does not always fully adjust the total population estimates to the known total population. This indicates that the goal of consistency with external data is not complete until the weights have gone through the calibration procedure. Särndal (2007) wrote "A desire to promote credibility in published statistics is an often-cited reason for demanding consistency." In other words, the potential for embarrassment due to a contradiction in known totals, or consistency of totals over time, is one of the reasons for the calibration process.

We want weights which satisfy the equation

$$\sum_{i \in s} w_i \mathbf{x}_i = \mathbf{x}.$$

where the sum of the sample estimates x_i using the calibration weights w_i are equal to the control totals x.

As indicated above, the weighting process in the CE Survey includes the following steps:

1. First, the base weights are determined by calculating the probability of initial selection for a household, based on the number of households selected in a

[6] Each $\mathbf{x}_i$ is a 35x1 vector of 0's and 1's indicating whether a respondent household is in a particular demographic group.

[7] $\mathbf{x}$ is another 35x1 vector with the Current Population Survey's known total number of households that are in those demographic groups.

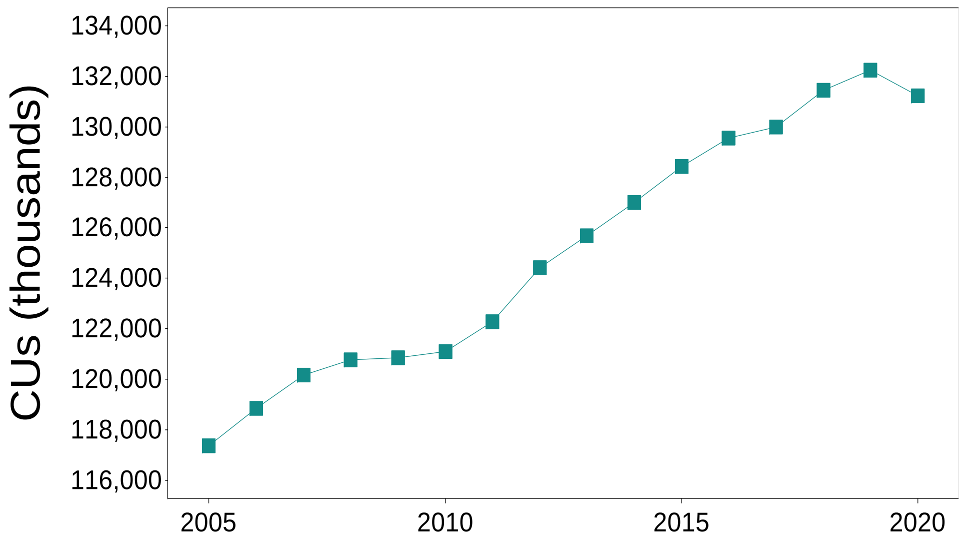
particular geographic area. Thus, the base weights are equal within a given metropolitan area for a given year.

2. Next, this weight is multiplied by a nonresponse adjustment factor, producing a CU-level weight, which we call NONINTWT.
3. Finally, this weight is multiplied by a calibration adjustment factor, which produces the calibration final weight, which we call FINLWT21.

Consistency with external data is an automatic consequence of calibration, and therefore not really the goal of calibration in the CE Survey. However, this is only the case for the 35 population counts to which the data are calibrated.

For example, the following table shows the weighted number of households (or "consumer units") in the United States, as published annually in the CE Tables:

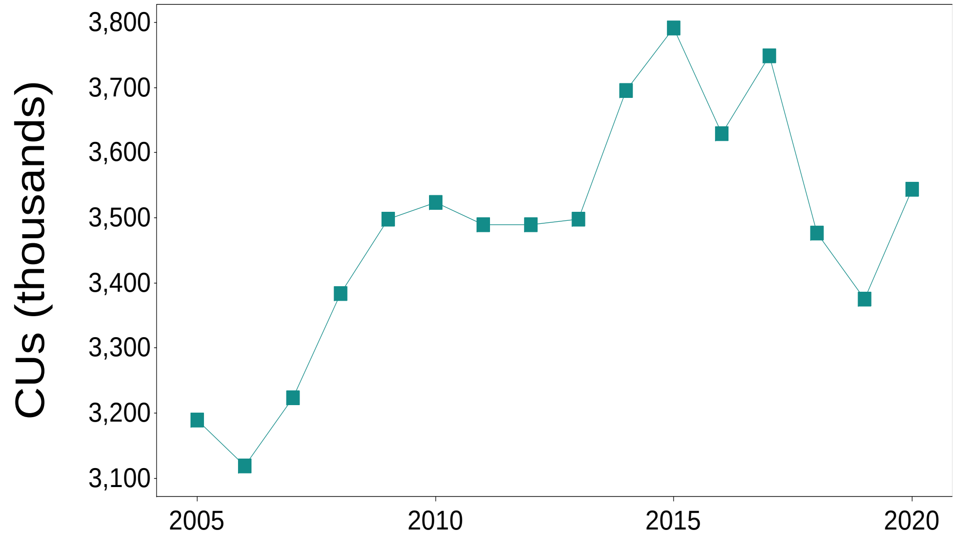
Figure 1: Weighted Number of Households in the CE Survey, 2005-2020



The figures shown in this graph match figures shown in other external sources, which is consistent due to the fact that total number of households in the country is one of the known population counts to which CE data are calibrated.

However, this can be contrasted with the less steady figures shown in the following graph, of values which are also published annually by the CE Survey, for the estimated number of households in the Chicago Metropolitan Area:

Figure 2: Weighted Number of Households in the Chicago Metropolitan Area in the CE Survey, 2005-2020



As can be seen from the above graph, there is a potential for embarrassment in the vacillating nature of these data, which would be much more steady were the CE Survey to calibrate to the total of, in this case, number of households in a given metropolitan area.

2.4 Calibration for Correcting the Sampling Weights to Account for an Incomplete Sampling Frame in the CE Survey

The calibration process takes the already-calculated nonresponse weights and adjusts them to what become calibration weights. Here as an example are the nationwide sum of the weights from the first quarter of 2021, in the CE Survey. The total sum of the weights is intended to represent the total population of households in the country (roughly 130 million):

BASEWT	NONINTWT	FINLWT21
57.3 million	121.2 million	132.8 million

As this example indicates, the average nonresponse adjustment factor is thus around 121.2/57.3 ≈ 2.1, and the average calibration adjustment factor is thus around 132.8/121.2 ≈ 1.1. While relatively small, the calibration adjustment factor is therefore instrumental in bringing our total for the weights up to the expected population total. Therefore, it helps complete the task begun by the nonresponse adjustment factor in adjusting for coverage errors, thus accounting for an incomplete sampling frame.

Sverchkov *et al* (2005) wrote that the same sampling frame is often used, as in the CE Survey, for years at a time, even though the frame changes over time, which can lead to biased estimates. When this drift in the sampling frame occurs, the calibration process can make a correction of this without the need for additional sampling.

However, while this was true about the CE Survey in 2005, now Census uses a "Master Address File" (MAF) to select households each year for the CE Survey. The MAF is updated every six months with data which come from the United States Postal Service. Therefore, this drift in the sampling frame no longer occurs, eliminating the need for correction via the calibration process.

Another way in which the sampling frame can be incomplete is due to the possibility that certain groups of the population are under-represented in the survey. To this end, the CE Survey is already looking at increasing sampling of lower-income groups, believed to be under-represented.

3. Conclusion

In conclusion, there are many ideas out there for what the goals of calibration should be, and how these goals can be applied to the CE Survey. The calibration process used by the CE Survey seems to be currently focused on minimizing variance, which might not be the best idea, particularly if a reduced variance estimation is only an underestimation of the true variance.

The nonresponse adjustment process, which occurs in the CE Survey prior to calibration, has been shown to work – that is, a recent study found no significant nonresponse bias in the published CE data. However, the nonresponse adjustment process also seeks to account for the entire sampling frame, and it has been shown that the calibration process furthers that goal by adjusting for coverage errors in the sampling frame. This should be the main goal of calibration in the CE Survey moving forward.

4. Acknowledgements

The views expressed in this paper are those of the author and do not necessarily reflect the policies of the U.S. Bureau of Labor Statistics.

References

Andersson, P. (2019). "'Optimal' calibration weights under unit nonresponse in survey sampling," *Survey Methodology* (Statistics Canada), Vol. 45, No. 3, pp. 533-542.

Bruce, S. and Elliot, D. (1998). "The Use of Calibration Weighting on Large Government Surveys," *New Methods for Survey Research: Proceedings of the International Conference*, pp. 95-106.

Davies, G. (2018). "Examination of approaches to calibration in survey sampling," doctoral thesis, Cardiff University.

Deming, W. E. and Stephan, Frederick F. (1940). "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known." *The Annals of Mathematical Statistics*, Vol. 11, No. 4, pp. 427-444.

Dever, J. (2008). "Sampling weight calibration with estimated control totals," doctoral dissertation directed by Professor Richard Valliant, JPSM.

Deville, J.-C. and Särndal, C.-E. (1992). "Calibration Estimators in Survey Sampling," *Journal of the American Statistical Association*, Vol. 87, No. 418, pp. 376-382.

Estevao, V., Hidroglou, M. A., and Särndal, C.-E. (1995). "Methodological Principles for a Generalized Estimation System at Statistics Canada," *Journal of Official Statistics*, Vol. 11, No. 2, pp. 181-204.

Estevao, V. and Särndal, C.-E. (2000). "A Functional Form Approach to Calibration," *Journal of Official Statistics*, Vol. 16, No. 4, pp. 379-399.

Folsom, R. E. and Singh, A. C. (2000). "The Generalized Exponential Model for Sampling Weight Calibration for Extreme Values, Nonresponse, and Poststratification," *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 598-603.

Jayasuriya, B. R. and Valliant, R. (1996). "An Application of Restricted Regression Estimation in a Household Survey," *Survey Methodology*, Vol. 22, No. 2, pp. 127-137.

Kott, P.S. (2006). "Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors," *Survey Methodology*, Vol. 32, No. 2, pp. 133-142.

Kott, P.S. (2012). "An Introduction to Calibration Weighting for Establishment Surveys," *American Statistical Association.* https://ww2.amstat.org/meetings/ices/2012/papers/302286.pdf

Lundstrom, S., and Särndal, C.-E. (1999). "Calibration as a Standard Method for Treatment of Nonresponse," *Journal of Official Statistics*, Vol. 15, No. 2, pp. 305-327.

Särndal, C.-E. (2007). "The Calibration Approach in Survey Theory and Practice," *Survey Methodology*, Vol. 33, No. 2, pp. 99-119.

Steinberg, B., Krieger, S., McBride, B., Nix, B., Sverchkov, M., and Yang, D. (forthcoming). "A Nonresponse Bias Study of the Consumer Expenditure Survey for the Ten-Year Period 2007-2016." U.S. Bureau of Labor Statistics.

Sverchkov, M., Dorfman, A., Ernst, L., Moerhle, T., Paben, S., and Ponikowski, C. (2005). "On Calibration and Non-response Adjustment for National Compensation Survey." U.S. Bureau of Labor Statistics.

Zieschang, K. D. (1990). "Sample Weighting Methods and Estimation of Totals in the Consumer Expenditure Survey," *Journal of the American Statistical Association*, Vol. 85, No. 412, pp. 986-1001.