Improving Precision by Calculating Estimates during the Calibration Process

Marcus E. Berzofsky[1], Bonnie E. Shook-Sa[1], G. Lance Couzens[1]
Philip S. Kott[1]
[1]RTI International, 3040 Cornwallis Rd, Research Triangle Park, NC 27709

Abstract

Analysts typically employ calibrated weights to improve their inferences. Calibration weighting can adjust for differing probabilities of response or coverage among the elements in the sampled population or it can reduce the impact of purely random error sources. Final weights may have undergone several calibration-adjustment steps. Each adjustment can increase the variability of the weights, which results in a decrease in the precision as conventionally measured. Relatively new software such as SUDAAN's WTADJUST procedure and R allow analysts to produce more accurate precision measures by calculating estimates during a calibration-weighting step rather than after. This means that calibrated weights need no longer be treated as if they were the original design weights (i.e., the inverse-selection-probability weights) in standard-error estimation. Using recalibrated sub-national estimates of victimization rates from the National Crime Victimization Survey, a nationally representative survey of persons in households, sponsored by the Bureau of Justice Statistics, we show the extent to which this improved measure of precision can reduce estimated standard errors.

Key Words: Post-stratification, Calibration, Precision, National Crime Victimization Survey (NCVS)

1. Introduction

1.1 Background

Weight calibration is employed by survey practitioners to improve inference to the target population by reducing bias in estimates. Typically, by the time a final analytic file is created, the initial design-based weights (i.e., the inverse probability of selection) have gone through several adjustments. The adjustments that are typically employed account for differing probabilities of nonresponse and undercoverage (Oh & Scheuren, 1983; Holt & Smith, 1979). These adjustments can also reduce the impact of random error sources through knowing the underlying nonresponse or population model (Little & Vartivarian, 2006).

Weight calibration adjustments have counteracting effects on the estimated precision of resulting point estimates. While these weight adjustments are critical to reducing bias, they typically have the offsetting effect of decreasing the precision as conventionally measured by increasing the variability in the weights. However, if the calibration model fits the data well and estimates are produced at the time of calibration, calibration can lead to a

reduction in the estimated standard errors of resulting point estimates. Therefore, statisticians need to balance the level by which bias can be reduced with the impact on precision when developing analytic weights. It is well documented that calibration results in an increase in estimated variances due to increased variation in analytic weights (Little & Vartivarian, 2006). However, the potential reduction in variance estimates resulting from calculating estimates at the time of calibration has not been quantified using real-world survey data.

In this paper we compare two approaches to producing the final set of calibrated weights and survey estimates to determine if one reduces the estimated variance of the estimate in a meaningful way. First, a *single-step approach*, calculates the estimates during the final calibration step. Second, a *two-step approach*, calculates the estimates in a separate procedure after the final calibration step has been conducted. Differences between the two approaches provide an estimate of the reduction in estimated variances due to the calibration model.

1.2 Estimation methods

Each of the two estimation methods should produce identical point estimates, though the manner in which available information is utilized differs between them. In this section we describe each method along with the advantages and disadvantages of each.

1.2.1 Single-step approach

Relatively new software (e.g., SUDAAN 11, R) allows analysts to produce estimates during the final calibration step rather than in a separate step after the calibration weights are produced. Under this approach, procedures such as SUDAAN's WTADJUST, will produce descriptive statistics along with the calibrated weights.

As shown in ***Table 1***, this approach has the advantage of yielding lower and more accurate standard error estimates because it takes into account information from outside the sample while estimating the variance. However, in order to take advantage of this benefit, one must reproduce the calibration each time an analysis is conducted. In order to do this one must have all the information needed for the final calibration being conducted (e.g., control totals for a coverage adjustment and pre-calibrated weights) available to them. If this information is not available or if the model being used (e.g., any interactions in addition to main effect terms) is not known, then the results produced by one analyst may not be reproducible by another analyst.

Table 1: Advantages and Disadvantages of the Single-Step Approach

Advantages	Disadvantages
• Should yield more accurate, lower variance estimates	• Requires reproduction of calibration for each analysis • Variance estimates are not easily reproducible by other researchers • Only leverages precision increases from the last calibration step; cannot incorporate prior calibration steps

1.2.2 Two-step approach

The two-step approach is the method typically used by survey analysts. Under this approach weights and estimation are conducted in two distinct steps. In the first step, the final calibration is conducted – using a procedure such as WTADJUST – and the weights are appended to the survey data. In the second step, in a separate procedure, such as SUDAAN's DESCRIPT, the final weights are used to produce the survey estimates.

As detailed in ***Table 2***, the two-step approach allows for easily reproducible variance estimates because the survey weights are fixed on the analysis dataset. However, because these weights are treated as fixed and the estimation process cannot account for information outside of the sample, the estimate of the variance is likely to be artificially higher.

Table 2: Advantages and Disadvantages of the Single-Step Approach

Advantages	Disadvantages
• With access to the recalibrated weights, analysts can easily reproduce the variance estimates	• Variance estimates are artificially higher

2. Case Study: The National Crime Victimization Survey

The National Crime Victimization Survey (NCVS), conducted since 1973 and sponsored by the Bureau of Justice Statistics (BJS), is a nationally representative sample of approximately 50,000 households and 75,000 persons interviewed twice per year. The survey provides estimates of the frequency and characteristics of non-fatal crime victimization in the United States.

The survey is designed to produce national estimates; however, BJS recognizes the importance of subnational estimates (e.g., state and metropolitan statistical area (MSA)). As such, BJS is interested in assessing a multi-prong approach to producing subnational estimates. Namely, (1) direct estimates, and (2) model-based small area estimates.

In this paper, we focus on producing direct estimates. In order to produce direct estimates with reliable precision, only areas in self-representing primary sampling units were included in this analysis. These areas included seven states and twenty-two MSAs with sample sizes large enough to produce estimates with reasonable precision. ***Figure 1*** and ***Figure 2*** illustrate the seven states and twenty-two MSAs included in the analysis.

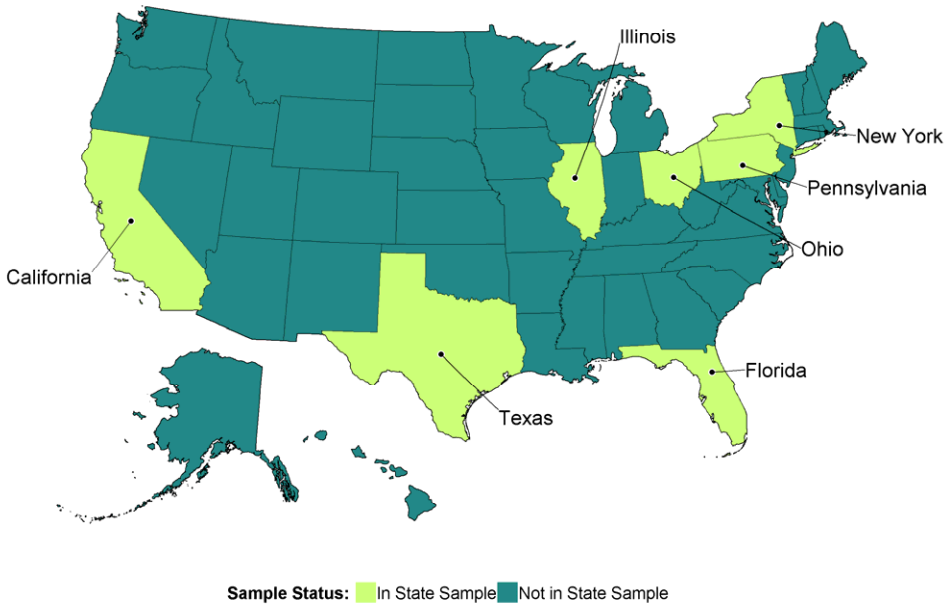

Figure 1: States included in subnational estimates analysis

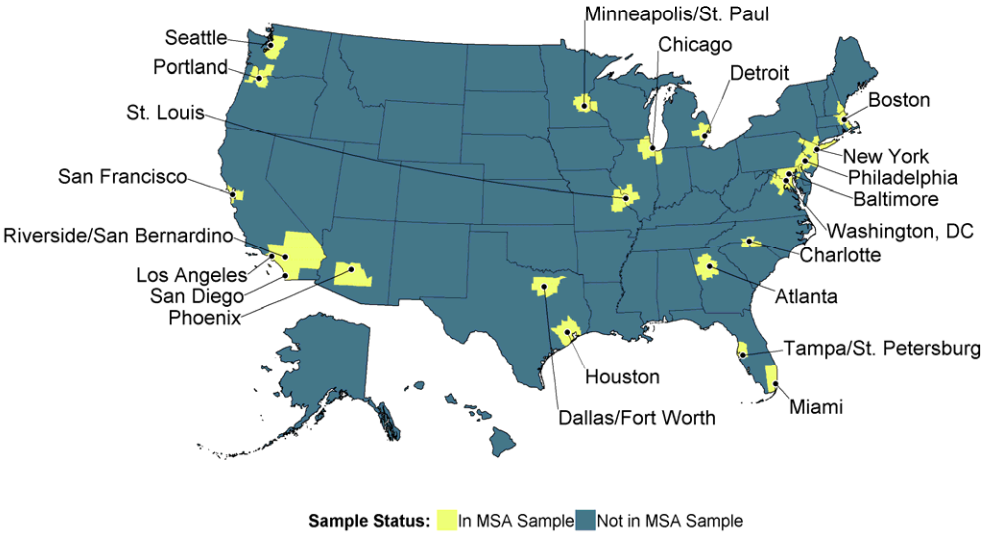

Figure 2: MSAs included in subnational analysis

Within each area, BJS is interested in producing unbiased estimates for overall crime types[1], crimes reported to the police, and total violent and property crimes by demographic

[1]Overall crime types include total violent, serious violent, rape and sexual assault, aggravated assault, robbery, simple assault, total property, household burglary, theft, and motor vehicle theft.

characteristics with precision that is considered *reliable*. BJS defines a reliable estimate as one with an unweighted sample size of more than 10 and a relative standard error (RSE) less than 50%.

Because the NCVS weights are calibrated for national estimates, an assessment comparing the weighted NCVS distributions to the distributions in the American Community Survey ACS) across key demographic characteristics[2] was conducted. Using data from 2008 – 2012, the 2012 1-year, 2010-2012 pooled 3-year, and 2008-2012 pooled 5-year distributions were compared. The analysis found that that the NCVS weighted distributions differed significantly from the ACS distributions for several of the characteristics – especially for population totals – in each subnational area. Thus, the national weights were re-calibrated to the population totals in each area to reduce potential bias in point estimates (Shook-Sa, et. al., forthcoming).

Re-calibrating the weights for each subnational area was expected to decrease the estimated precision in resulting estimates. However, because of limited sample sizes in subnational areas it was desired that the selected estimation approach minimize the impact on precision estimates to the extent possible. Therefore, in order to assess the estimation process that would minimize the estimated variance for the direct subnational estimates, we compared the two estimation methods to answer the following research questions:

1. Does the single-step method produce smaller standard errors than the two-step approach?
2. If the single-step method improves precision, does the reduction in the standard errors under the single-step approach increase the number of estimates that are deemed reliable?

3. Methods

Within each subnational area, estimates were produced using both calibration methods.[3] For each area, a marginal calibration model was fit using ACS population totals and all comparable characteristics across both surveys. When a model for an area did not converge, a model reduction algorithm was implemented to produce a convergent model. Once a suitable calibration model was found for an area, estimates were produced under each of the estimation methods.

Once estimates were produced under each method, the approaches described in the following sections were used to assess each research question.

[2] Characteristics compared include the following person-level characteristics: gender, age category, race/ethnicity, percent federal poverty level (FPL) of the household, household tenure, educational attainment, marital status, and employment; and the following household-level characteristics: age of householder, race/ethnicity of householder, percent FPL of the household, household tenure, educational attainment of householder, number of housing units in structure, and number of motor vehicles.

[3] Because the lowest level of geography available on the NCVS public use file is Census region (due to disclosure concerns), the re-calibration and estimation of the NCVS estimates needed to be conducted within a Census Research Data Center (RDC).

3.1 Point Estimates

Prior to assessing the impact the calibration methods have on precision, we verified that the two methods produced the same point estimates. Verifying the point estimates ensures that the two methods are working as expected. In order to compare point estimates, the estimates for all outcomes were directly compared. This included 9,657 estimates across all areas, year groups and estimate types.

3.2 Reduction in estimated variance

To determine if, and by how much, the single-step method reduced the estimated variances compared to the two-step method the percent change in the estimated standard errors was computed for all estimates in each area, year group, and estimate type. The distribution of the change was assessed to understand the magnitude by(?) which the variance estimates changed.

3.3 Change in the number of reliable estimates

Once the standard errors are computed for each estimate, the relative standard errors were computed. The proportion of estimates deemed unreliable under each method were compared.

4. Results

4.1 Verification of Point Estimates

Figure 3 presents the comparison of the 9,657 point estimates for 1-year estimates in states and MSAs. In the figures, points along the 45-degree line are estimates that are exactly the same under each method. As expected, it is clear that all points fall along the 45-degree line, confirming that point estimates are not impacted by the estimation method. This finding was true for 3-year and 5-year estimates as well.

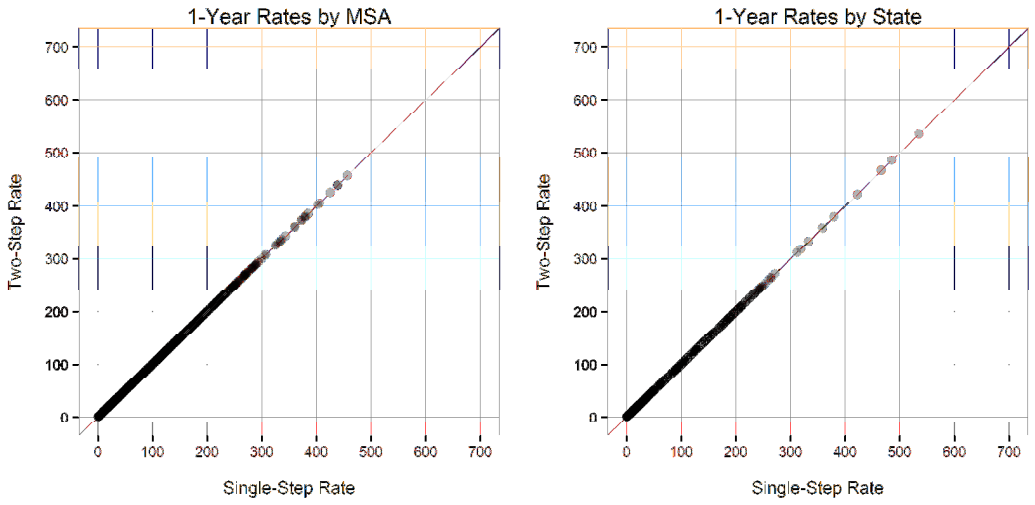

Figure 3: Comparison of 1-year point estimates by single-step and two-step methods within MSAs and States

4.2 Reduction in estimated variance

Figure 4 presents the distribution of the percent change in the standard errors for estimates under each approach by area type and year grouping. As the figure shows, the single-step method overwhelmingly produces standard errors that are smaller than the two-step method. Among MSAs, the single-step method produces smaller standard errors in 85.0% to 87.3% of estimates depending on the year group. Among states, the single-step method produces smaller standard errors in 90.2% to 92.5% of estimates depending on the year group. However, as the figure shows, the large majority of standard errors (84.3%) are reduced by 5% or less.

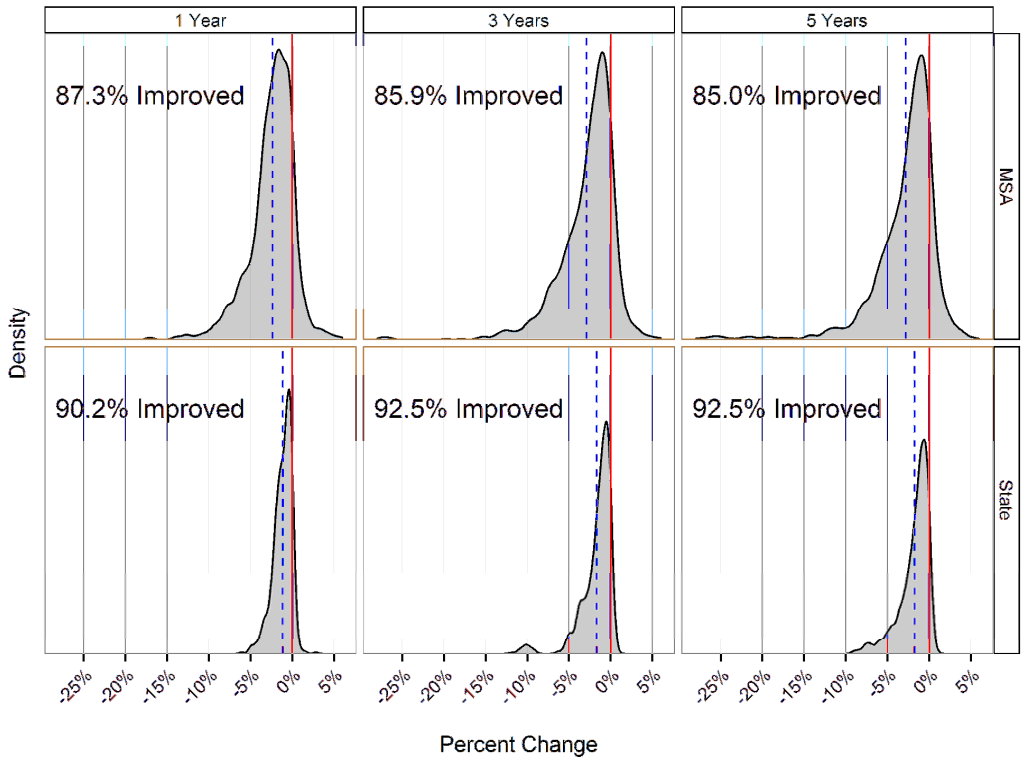

Figure 4: Distribution of the percent change in standard errors under the two estimation methods by area type and year group (means denoted by dashed blue lines).

4.3 Change in the number of reliable estimates

Figure 5 presents the number of estimates for which the reliability remained unchanged under both methods, was improved under the single-step method, and was improved under the two-step method. As can be seen, while the single-step method improved the precision for more estimates than the two-step method, the reliability for the vast majority of estimates did not change. For example, under a 50% RSE reliability rule for all area and year groups, 29 (out of the possible 9,246) additional estimates were reliable under the single-step approach compared to the two-step approach. No additional estimates were reliable under the two-step approach. This was true for a reliability criteria of an RSE less than 50% and an RSE less than 30%[4].

[4] While not the criteria used by BJS, an RSE of 30% or less is the criteria used by other large federal surveys to classify an estimate as reliable.

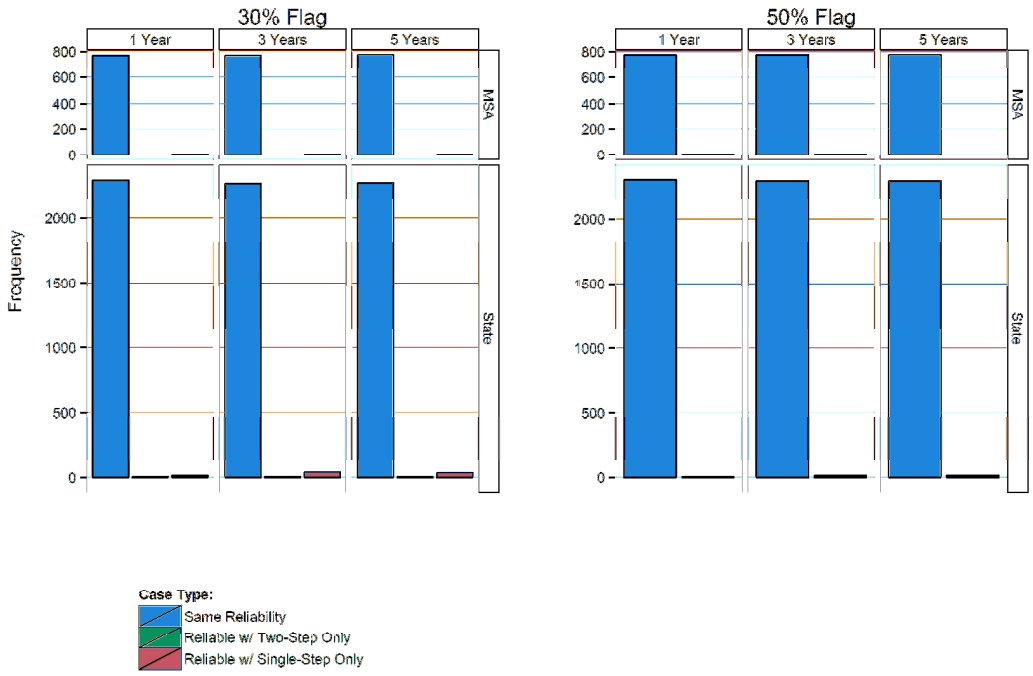

Figure 5: Number of estimates whose reliability status was the same under the single-step and two-step approach, improved under the single-step approach, and improved under the two-step approach by area type and RSE threshold.

5. Conclusions

The need to produce unbiased estimates without overly inflating variance estimates is critical when producing estimates with multiple calibration adjustments and where the sample size may be relatively small. This is the case for the NCVS when producing direct estimates in subnational areas. In order to minimize the estimated variances, two estimation methods where compared – a single-step approach and a two-step approach. Our analysis of these two methods using the NCVS data found that while the single-step method does reduce the estimated variance it does not do so in a meaningful way. In other words, the number of estimates that are reliable under the single-step approach, but not under the two-step approach, are negligible.

These findings have two main implications. First, while the magnitude of precision gains is small, the single-step approach does lead to smaller variance estimates for the large majority of estimates. Therefore, if able, it is recommended to use the single-step approach. Second, for analysts that do not have access to the calibration totals or the final calibration model, the two-step approach – which is the approach used by most analysts – does not, at least for the NCVS data, detrimentally impact precision. We expect that these findings would hold for other datasets, but further analysis should be conducted to verify.

Acknowledgements

The authors would like to thank BJS for sponsoring this research and comments provided by Mike Planty and Lynn Langton. However, we would like to note that the views

expressed in this paper are those of the authors only and do not reflect the views or position of BJS or the Department of Justice.

References

Holt, D., and Smith, T.M.F. (1979). Post stratification. *Journal of the Royal Statistical*, *A*, 142, 33-46.

Little, J.L. and Vartivarian, S. (2006). Does weighting for nonresponse increase variance of survey means? *Survey Methodology,* Vol. 31 (2), 161 – 168.

Oh, H.L., and Scheuren, F.S. (1983). Weighting adjustments for unit nonresponse. In *Incomplete Data in Sample Surveys*, 2, Theory and Bibliographies, (Eds. W.G. Madow, I. Olkin and D.B. Rubin), Academic Press, New York, 143-184.

Shook-Sa, B., Berzofsky, M.E., Couzens, L., Moore, A.., and Lee, P. (2015). *Assessing the Coverage and Reliability of NCVS Sample in the Largest States, MSAs, and Cities.* Prepared for the Bureau of Justice Statistics, Washington, DC.