Survey Research Methods (2017)
Vol. 11, No. 4, pp. 405-414
doi:10.18148/srm/2017.v11i4.6789

ISSN 1864-3361
http://www.surveymethods.org

Assessing Alternative Precision Measures when Adjusting for Conditional Bias at the Subnational Level through Calibration Weighting

Bonnie E. Shook-Sa
Department of Biostatistics
University of North Carolina at Chapel Hill

Phillip S. Kott
RTI International

Marcus E. Berzofsky
RTI International

G. Lance Couzens
RTI International

Andrew Moore
RTI International

Philip Lee
RTI International

Lynn Langton*
Bureau of Justice Statistics

Michael Planty
RTI International

Calibration weighting improves inference by adjusting for observed differences between the realized sample and the population. Unfortunately, a commonly-used linearization-based variance estimator often does not account for the increased efficiency provided by the calibration process. As a result, precision estimates based on calibrated weights can be artificially high. Using a corrected linearization-based variance estimator that was recently made easier to compute allows analysts to utilize calibration-weighting techniques while producing more accurate precision estimates.

We use calibration weighting to produce more reliable subnational estimates and assess the differences in point estimates resulting from these weight adjustments in the National Crime Victimization Survey, a nationally representative survey designed to calculate victimization rates solely at the national level. We then assess the estimated precision of these point estimates using a conventionally implemented linearization-based variance estimator and the corrected variance estimator. We find that the calibration adjustments mostly reduced the standard errors in subnational estimates but to successfully measure the reduction required using the corrected variance estimator.

Keywords: variance estimation; calibration weighting; National Crime Victimization Survey

1 Background

Survey practitioners often employ calibration weighting to reduce bias in survey estimates subject to nonresponse and undercoverage (Holt & Smith, 1979; Oh & Scheuren, 1983). When covariates in the linear model underlying calibration weighting are related to survey outcomes, calibration can also lead to a reduction in the standard errors of resulting point estimates (Little & Vartivarian, 2006). Sadly, conventionally implemented linearization-based variance estimation does not capture the full impact of calibration weighting and can lead to upwardly biased variance estimates (Research Triangle Institute, 2012, p. 1027). Often this results in an oversimplified view of a bias/variance tradeoff associated with calibration (Little & Vartivarian, 2006).

With the use of appropriate variance estimation techniques, such as replication, calibration weighting can be implemented to reduce conditional bias in estimates while avoiding artificially inflated variance estimates (see, for example Valliant, Dever, & Kreuter, 2013, p. 421). Unfortunately, calibration can sometimes fail in replicate samples when calibration and replicate weights are forced to be positive, rendering variance estimators using such a replication method less than ideal. Use of conventionally implemented linearization techniques to produce variance estimates for surveys that utilize calibrated weights remains common (e. g. United States Department of Justice, Office of Justice Programs, & Bureau of Justice Statistics, 2014; Ward, Clarke, Nugent, & Schiller, 2016). For this reason, it is critical to understand and quantify differences in precision

* Views expressed in this paper are those of the author and do not represent the official views or position of the U.S. Department of Justice

Contact information: Bonnie E. Shook-Sa, DrPH Student, Department of Biostatistics, University of North Carolina at Chapel Hill, 170 Rosenau Hall, CB #7400, 135 Dauer Dr., Chapel Hill, NC 27599, (E-mail: bshooksa@live.unc.edu)

estimates based on conventionally implemented versus corrected variance estimation approaches that take into account the effects of calibration weighting on variance estimates. This paper quantifies differences in precision estimates using conventionally implemented linearization-based variance estimation and a corrected linearization-based method that appropriately takes into account efficiencies from calibration (Research Triangle Institute, 2012) using data from the National Crime Victimization Survey (NCVS).

The NCVS is sponsored by the U.S. Bureau of Justice Statistics (BJS) and produces estimates on the incidence and characteristics of criminal victimization in the United States. The NCVS is a nationally-representative sample of approximately 40,000 households and 75,000 persons per year. The NCVS is based on a stratified multi-stage cluster sample and is a rotating panel design. At the first stage, Primary Sampling Units (PSUs) consisting of counties, groups of counties, or large metropolitan areas are selected. PSUs are grouped into strata, with large PSUs included in the sample with certainty as self-representing PSUs and the remaining, non-self-representing PSUs grouped into strata based on geographic and demographic characteristics from the decennial Census. Within selected PSUs, a two-stage sample is selected, with (1) the selection of enumeration districts of approximately 750-1,500 persons and (2) segments of about four housing units each selected within each enumeration district (United States Department of Justice, Office of Justice Programs, & Bureau of Justice Statistics, 2010). All housing units within selected segments are sampled for the NCVS. Within selected housing units, a household respondent, which is an adult knowledgeable about the household, is selected to complete the household interview. The household interview contains questions about the household and criminal victimizations experienced by the household over the past 6 months, which serves as the reference period. All persons 12 and older within selected households are selected to complete the person-level interview, which contains questions about the person and criminal victimizations experienced by the person within the reference period. Households selected for the NCVS are interviewed once every six months for a three-year period, after which they are rotated out of the sample and new households are rotated into the sample.

Despite its large sample, the NCVS has historically been designed and weighted solely to produce national estimates. With this in mind, BJS is working to enhance the NCVS through the development of a subnational estimation program. One component of this program entails exploring whether the current NCVS sample in large states and metropolitan statistical areas (MSAs) can be used to produce reliable and valid estimates using direct estimation techniques when data are pooled across multiple years. These subnational crime statistics could then be used to better understand local crime patterns and long-term trends through the assessment of rolling pooled estimates.

Because the NCVS sample was designed and weighted for national estimation, controls are not in place to reduce conditional bias in state-level estimates caused, for example, by the random choice of primary sampling units (PSUs) selected without concern for the need to make subnational estimates. "Conditional bias" in this context refers to the contribution to the standard error of a subnational estimate under probability sampling theory caused by differences between the weighted sample and population of interest in control variables correlated to the variable being estimated. The term is used here in the spirit of Royall (1971), but more loosely than in his precise model-based sense in which the survey variable is assumed to be a random variable with an expectation equal to the same linear function of the calibration variables whether or not the population unit is in the sample (see Section 2.2 for the sense meant here).

We discuss the development of calibrated, area-specific subnational weights for the NCVS at the state and MSA level to reduce potential conditional bias in subnational estimates. We compare point estimates based on the national and subnational calibrated weights to demonstrate the differences when the national weights are applied at the subnational level. We then compare variance estimates based on the conventionally implemented and corrected linearization-based variance estimation methods to demonstrate how far off the conventional estimates can be. Finally, we examine the types of estimates for which the corrected variance estimator has the largest impact.

2 Methods

2.1 Motivation for subnational recalibration

BJS is interested in producing reliable, nearly unbiased estimates for key crime types in the seven largest states and 22 largest MSAs in the United States. "Nearly unbiased" here means that the squared bias is an asymptotically ignorable contributor to mean squared error when the number of sampled PSUs is large. Formally, the bias tends toward zero as the number of sampled PSU's grows arbitrarily large. These states and MSAs are the subnational areas that were determined to have sufficient sample sizes to support direct estimation with the current NCVS sample. See Figure 1 and Figure 2.

For estimates to be efficient, the sample within each subnational area must represent the target population. As noted previously, the NCVS was designed and weighted with the goal of producing nearly unbiased and relatively efficient national estimates. Stratification and allocation of the sample were implemented with the goal of selecting PSUs that are nationally representative. Although the subnational-level sample is nearly unbiased in expectation, an unfortunate

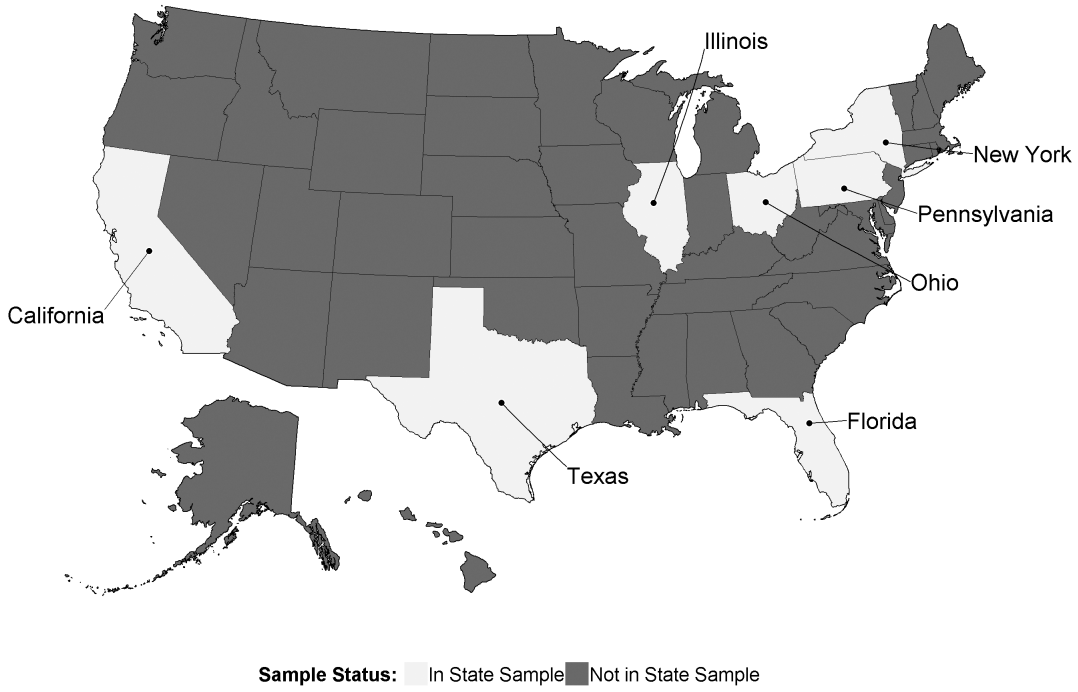

Figure 1. States included in subnational estimates analysis

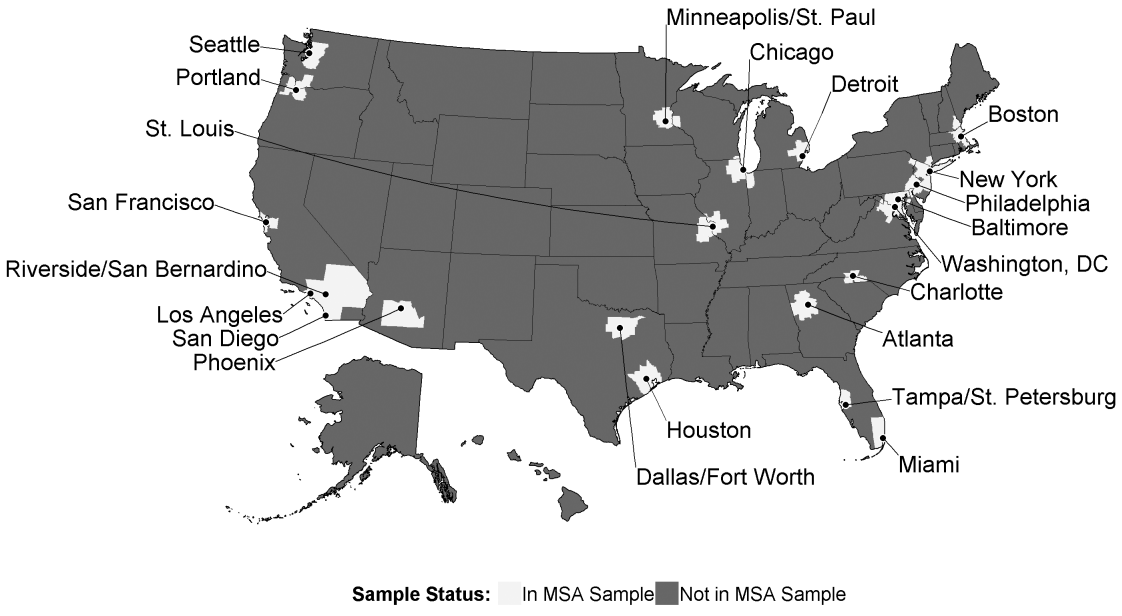

Figure 2. MSAs included in subnational estimation analysis

sample selection could lead to conditional bias at the subnational level. This is a source of additional standard error that can be corrected by recalibrating weights based on control totals at the subnational level.

2.2 Calibration weighting and comparisons of estimates based on recalibrated and national weights

Because of the limitations of the national weights, NCVS person and household weights were recalibrated to a set of population totals, or strong estimates of those totals from an external source, in each area to reduce potential conditional bias in subnational estimates. An exponential calibration model, $w_k = d_k \exp(\boldsymbol{x}_k^T \boldsymbol{g})$, was used, where k denotes a sample respondent, w the recalibrated weight, d the original weight before recalibration, $\boldsymbol{x}$ a vector of control variables, and $\boldsymbol{g}$ is chosen to satisfy a calibration equation: $\sum w_k \boldsymbol{x}_k = \boldsymbol{T}_x$, where the summation is over the sample in an area of interest (i. e., a subpopulation for which estimates are desired), and $\boldsymbol{T}_x$ is the vector of totals for the control variables in that area (Deville & Sãrndal, 1992; Folsom & Singh, 2000). This equality removes the conditional bias in a pre-calibrated estimator for a total in the area of interest, $t = \sum d_k y_k$, that results from the likely inequality given the realized sample between $\sum d_k \boldsymbol{x}_k$ and $\boldsymbol{T}_x$. Although $\sum d_k \boldsymbol{x}_k = \boldsymbol{T}_x$ on average across all possible samples, the equality is unlikely to hold when conditioned on a particular realized sample.

Royall (1971) assumes the y_k are random variables, each with mean $\boldsymbol{x}_k^T \beta$, so that $(\sum d_k \boldsymbol{x}_k - \boldsymbol{T}_x)^T \beta$ is literally the conditional bias given the sample of t as a predictor for the population total of the y_k. In contrast, the calibrated estimator for that total, $t_c = \sum w_k y_k$, is conditionally unbiased no matter the realized sample.

The exponential calibration model is an extension of raking to allow continuous control variables. Unlike with the linear calibration model, $w_k = d_k(1 + \boldsymbol{x}_k^T \boldsymbol{g})$, the exponential model cannot produce negative calibration weights (because $\exp(\theta)$ is always positive). Its use, however, increases the possibility that no $\boldsymbol{g}$ satisfies the calibration equation, something that can only occur under linear calibration when some of the components of $\boldsymbol{x}_k$ in the sample are linearly dependent on each other so that the matrix $\boldsymbol{X}$ with $\boldsymbol{x}_k$ in the kth column is not of full rank.

The household and person-level control variables in the calibration model for an NCVS area of interest are listed in Table 1. The target population totals for the control variables were derived from the American Community Survey (ACS) covering the same time period as the NCVS sample used in the evaluation (2010 – 2012). Because the lowest level of geography available on the NCVS public use file is Census region due to representativeness and disclosure concerns, the recalibration and estimation of the NCVS estimates needed to be conducted within a Census Bureau Research Data Center. Analyses of NCVS microdata must be completed at a Census Bureau Research Data Center after completing the appropriate application and certification processes.[1].

NCVS data were pooled across three years by stacking the 2010 to 2012 datasets and dividing the analytic weights by three. Using pooled data from 2010–2012 ensured sufficient sample sizes within areas of interest. Key NCVS estimates and their standard errors were then calculated using the two different variance estimators based on the subnational-specific weights (discussion to follow). In addition, the same set of estimates was calculated based on national NCVS weights with variances estimated in the conventionally implemented linearization-based manner for reasons to be discussed.

Recalibrated and national point estimates were compared to assess the differences that would result had national NCVS weights been applied at the subnational level. Within each subnational area, comparisons were made for overall crime types by characteristics of the crime (23 estimates), crimes reported to the police by characteristics of the crime (23 estimates), and estimates by demographic characteristics of the victim (65 estimates). This resulted in 710 total comparisons across the seven states and 1,734 total comparisons across the 22 MSAs. These estimate counts exclude estimates that were suppressed by the Census Bureau Data Review Board due to insufficient sample sizes.

2.3 Comparison of the conventionally implemented linearization-based and corrected variance estimators

With the conventionally implemented linearization-based variance estimator, the subnational-specific weights obtained through recalibration were merged back onto the analytic file. A separate analytic procedure was then applied to calculate design-consistent point and variance estimates within each subnational area. The VARGEN procedure in SUDAAN 11 (Research Triangle Institute 2012) was used to calculate point and variance estimates, but other "design-based" software would work equally well.

With the corrected variance estimator, variance estimates were calculated at the time of calibration for each subnational area. Relatively new software (e. g. WTADJUST in SUDAAN 11 as well as several routines in R) allows analysts to produce corrected variance estimates using equations during the final calibration step rather than in a separate step after the calibration weights have been produced. Because both the corrected and conventionally implemented linearization-based methods are based on the same set of calibration weights, these approaches yield identical point estimates. However, the corrected variance estimator produces variance estimates that are asymptotically unbiased under

[1] https://www.census.gov/ces/rdcresearch/index.html

Table 1
The household and person-level control variables in the calibration model for an NCVS area of interest

Person-Level Characteristics	Household-Level Characteristics
Gender	Age of householder
Age category	Race/ethnicity of householder
Race/ethnicity	Percent FPL of the household
Persons by percent federal poverty level (FPL) of the household	Household tenure
Persons by household tenure[a]	Educational attainment of householder
Educational attainment	Number of housing units in structure
Marital status	Number of motor vehicles
Employment	

[a] Persons by household tenure refers to the number of persons living in households that are owned or rented.

mild conditions. With replication, by contrast, calibration needs to be performed on the sample and every replicate sample before variances can be estimated.

The conventionally implemented linearization-based variance estimator simply treats the calibrated weights (w_k) as if they were design weights (d_k). This captures any increase in variance due to the calibrated weights being more variable than the design weights. The corrected approach replaces the survey variable (y_k) in the variance estimator for an estimated total, $t = \sum w_k y_k$ (with the summation over the sample in the area of interest), by the residual of its weighted regression on the vector of control variables ($e_k = y_k - \boldsymbol{x}_k^T \boldsymbol{b}$, $\boldsymbol{b} = (\sum w_j \boldsymbol{x}_j \boldsymbol{x}_j^T)^{-1} \sum w_j \boldsymbol{x}_j y_j$). See, for example, (Kott, 2006) (the w_j in $\boldsymbol{b}$ are the first derivatives of $w_j = d_j \exp(\theta_j)$ with respect to θ_j). Therefore, the corrected estimation approach has the advantage of capturing the full impact of calibration weighting on variance estimation.

Mathematically, the linearization variance estimator of a rate $r = \frac{\sum w_k y_k}{\sum w_k}$, where (again) the summations defining r are over the sample in an area of interest, is

$$v(r) = \frac{\sum_{h=1}^{H} \frac{n_h}{n_h - 1} \left[\sum_{j=1}^{n_h} \left(\sum_{k \in S_{hj}} w_k e_k \right)^2 - \frac{\left(\sum_{j=1}^{n_h} \sum_{k \in S_{hj}} w_k e_k \right)^2}{n_h} \right]}{\left(\sum \sum \sum w_k \right)^2}. \quad (1)$$

In equation (1), H is the number of variance strata containing individuals in the area of interest, n_h is the number of variance PSUs in variance stratum, and S_{hj} is the sample of individuals in variance PSU j of variance stratum h. In the conventional implementation, $e_k = y_k - r$, while in the corrected variance estimator $e_k = y_k - \boldsymbol{x}_k^T (\sum w_j \boldsymbol{x}_j \boldsymbol{x}_j^T)^{-1} \sum w_j \boldsymbol{x}_j y_j$ if the vector of control variables contains a constant or the equivalent; otherwise, it is $e_k = (y_k - r) - \boldsymbol{x}_k^T (\sum w_j \boldsymbol{x}_j \boldsymbol{x}_j^T)^{-1} \sum w_j \boldsymbol{x}_j [y_j - r]$. This assumes that the variance PSU j in h is composed entirely of individuals in the area of interest. The name "corrected" variance estimator derived from using a corrected form of e_k in equation (1).

In order to have enough degrees of freedom for MSA estimates, technically defined as the number of variance PSUs minus the number of variance strata, we treated census tracts as the variance PSUs when computing estimated variances using equation (1), while the variance strata were the design strata and self-representing design PSUs. Census tracts were subsets of the design PSUs and were fully contained in an area of interest. This treatment would result in the underestimation of variances when respondent values are correlated across census tracts but within the actual design PSUs. Note that calibration weighting may remove some across-tract-but-within-PSU correlation. Any such underestimation would affect the conventionally implemented and corrected linearization-based variance estimators similarly.

The variances for estimates calculated with national weights were computed using the conventionally implemented linearization-based method. The conventionally implemented linearization-based estimator was used for the national weights because the pre-calibration weights and national control totals were not available to evaluate the corrected estimator based on the national weights. To make this comparison fairer, census tracts were again treated as the variance PSUs when estimating variances.

We had two sets of point estimates, either the national or recalibrated weights, and three sets of variance estimates, the conventionally implemented linearization-based variance estimator for both sets of point estimates or the corrected variance estimator for the recalibrated estimates, from which we computed relative standard errors (RSE). The RSE is defined as the estimated standard error of a point estimate divided by the point estimate itself expressed in percentage terms. In addition to comparing the three RSEs, we determined the percentage of estimates that would be flagged as unreliable based on a percent RSE of 30 percent. Although several large national surveys use an RSE criteria of 30 percent for the flagging of unreliable estimates, the NCVS currently uses a flagging criteria of 50 percent (Williams et al., 2015). Finally, an examination was conducted to assess the types of estimates for which the corrected method leads to the largest differences in variance estimates compared to the conventionally implemented linearization-based method of estima-

tion.

3 Results

3.1 Comparison of recalibrated to nationally-weighted estimates

Weight recalibration led to significant differences in estimated victimization rates within both states and MSAs (Figure 3). Recalibration reduced the victimization rates for 66 percent of state-level estimates and 60 percent of MSA-level estimates. Victimization rates likely tended to be higher with the national weights than the recalibrated weights because a high proportion of the subnational samples was located in urban areas which tend to have higher victimization rates than rural areas. The recalibration process corrected the weights to match population gold standards by urbanicity, which resulted in lower victimization rates. Figure 3 shows, however, that a sizeable proportion of estimates calculated with the recalibrated weights are higher than those calculated with the national weights.

Despite the large proportion of estimates that shifted as a result of recalibration, the magnitude of the shifts tended to be small. As shown in Table 2, the median change in victimization rates (per 1,000 persons or households) was 0.8 for states and 1.4 for MSAs. Recalibration did lead to some substantively important shifts in estimates, however. The 90^{th} percentiles for the absolute change in victimization rates due to recalibration were 6.7 and 11.5 for states and MSAs, respectively.

3.2 Comparison of conventionally implemented linearization-based and corrected variance estimators

The observed differences in estimates produced with recalibrated and national weights motivates the need for recalibration. As previously discussed, however, such recalibration can lead to an increase in variance estimates when the estimation procedure does not appropriately account for efficiencies in the calibration model. Such a loss in precision is particularly problematic in subnational areas that contain limited sample sizes to support reliable estimates.

The effects of ignoring the impact of calibration weighting on variance estimates versus appropriately accounting for the efficiencies from recalibration in NCVS subnational estimates are demonstrated in Figure 4. When the conventionally implemented linearization-based variance estimator is used, recalibration leads to an increase in the percent RSE (i. e., a reduction in precision) over the nationally-calibrated method for the majority of estimates. In contrast to that, when the corrected variance estimator was used, the majority of estimates were revealed to be slightly more precise under recalibration than under national weighting. For states, 61 percent of estimates based on the conventionally implemented linearization-based variance estimation method had increases in RSEs resulting from calibration compared to 45 percent of estimates using the corrected variance-estimation method. The effect was slightly more pronounced in MSAs, as 68 percent of conventionally implemented linearization-based variance estimates and 48 percent of corrected variance estimates had higher RSEs following recalibration. By appropriately taking into account efficiencies in the calibration models, the corrected method better preserves the precision of subnational estimates compared to the conventionally implemented linearization-based method.

Comparisons were also made between the percentage of estimates flagged as unreliable based on having 10 or fewer sample cases or a percent RSE greater than 30 percent. These minimum reporting standards, while somewhat arbitrary, were adopted by the Bureau of Justice after review of the literature and other federal agency standards balanced with the demand for reporting substantive findings to key stakeholders. Because estimates calculated with the national weights and both calibration estimation methods are based on the same sample of cases, any differences in estimate flagging are due solely to differences in estimated precision (i. e. percent RSE). Figure 5 compares the percentage of estimates flagged as unreliable for the nationally-weighted, conventionally implemented linearization-based recalibrated, and corrected recalibrated methods. The percentage of estimates flagged as unreliable is similar across the three methods. For both states and MSAs, fewer estimates are flagged as unreliable when the corrected recalibrated estimation method is used than the other two methods, but these differences are slight. The conventionally implemented linearization-based recalibration method leads to the most estimates being flagged as unreliable. Both comparisons show that recalibration tended to reduce variance estimates within these subnational areas when the efficiencies in the calibration weighting were appropriately accounted for, but increased variance estimates when those efficiencies were not taken into account. These comparisons clearly demonstrate the common misconception that calibration weighting may remove bias but it increases variance discussed by Little and Vartivarian (2006).

The corrected recalibrated estimation method provides clear gains in variance estimates over the conventionally implemented linearization-based method. Nevertheless, the magnitude of the differences in precision estimates is small for the majority of estimates (Table 3). The median difference in RSEs between the conventionally implemented linearization-based and corrected methods is only 0.2 percentage points for states and 0.4 percentage points for MSAs. For some estimates the magnitude is much larger, however. The maximum difference in RSEs is 6.4 percentage points for states and 26.5 percentage points for MSAs.

As a final comparison between the two variance-

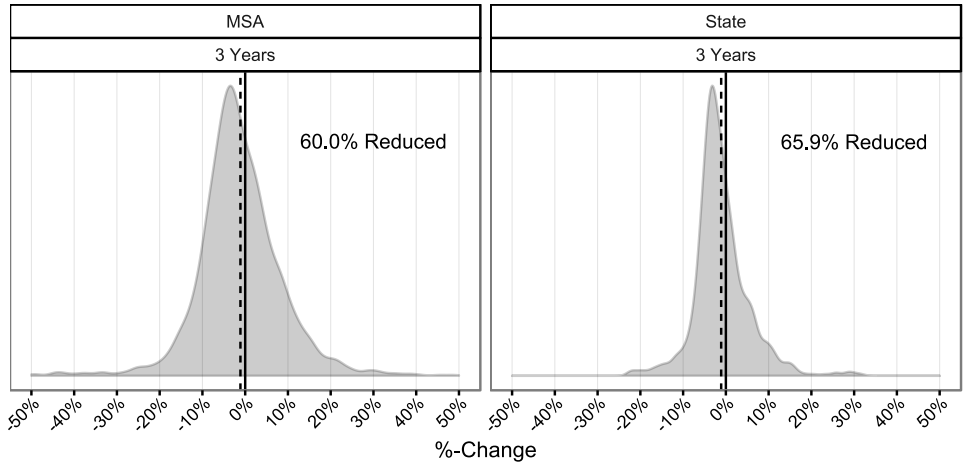

Figure 3. Percent change in NCVS victimization rates resulting from recalibration by subnational area type

Note: Mean is denoted by dotted line and 0% change is denoted by solid line. State distribution is based on 710 estimates and MSA distribution is based on 1,734 estimates.

Table 2
Distribution of absolute changes in victimization rates resulting from recalibration by subnational area type

Area Type	Minimum	10th Percentile	25th Percentile	Median	75th Percentile	90th Percentile	Maximum
States	0.0	0.1	0.2	0.8	3.3	6.7	232.3
MSAs	0.0	0.1	0.5	1.4	4.8	11.5	100.5

Note: MSAs=Metropolitan Statistical Areas
Differences were calculated by taking the absolute value of the estimate based on the recalibrated weights minus the estimate based on the national weights. Estimates are per 1,000 persons (personal crimes) or households (property crimes).

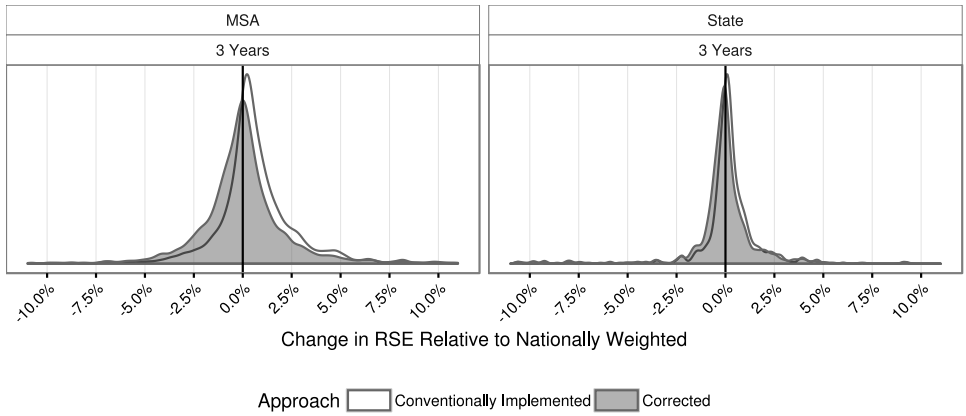

Figure 4. Absolute change in percent RSE for key estimates from nationally weighted to subnational specific using corrected and conventionally implemented linearization-based variance estimators by subnational area type

Note: 0% change is denoted by solid line. State distribution is based on 710 estimates and MSA distribution is based on 1,734 estimates.

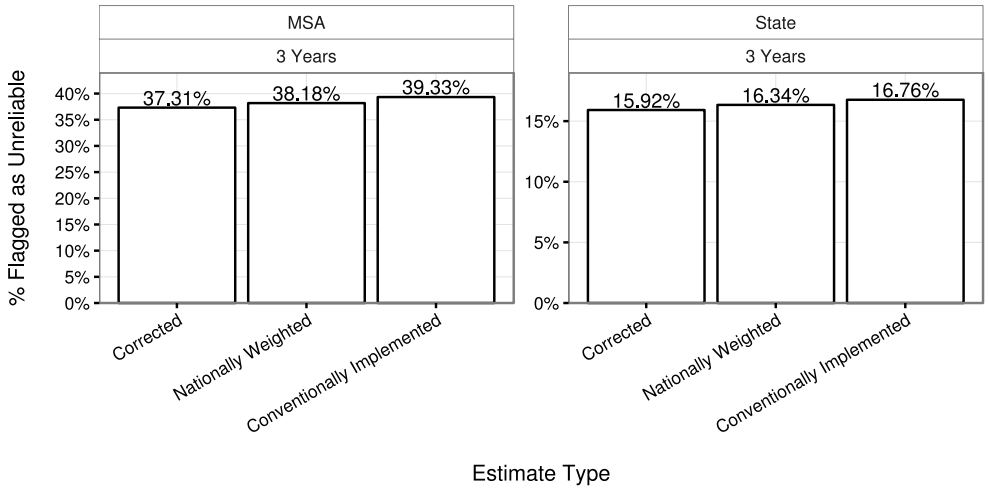

Figure 5. Percentage of estimates flagged as unreliable for nationally weighted, conventionally implemented linearization-based recalibrated estimation, and corrected recalibrated estimation methods by subnational area type

Note: Estimates are flagged as unreliable if they are based on an unweighted sample count of 10 or less or have a percent RSE greater than 30 percent. State distribution is based on 710 estimates and MSA distribution is based on 1,734 estimates.

Table 3
Distribution of difference in percent RSE between conventionally implemented linearization-based and corrected variance methods by subnational area type

Area Type	Minimum	10th Percentile	25th Percentile	Median	75th Percentile	90th Percentile	Maximum
States (in %)	−0.3	0.0	0.1	0.2	0.3	0.6	6.4
MSAs	−2.6	−0.1	0.1	0.4	1.0	1.8	26.5

Note: RSE = Relative Standard Error; MSAs=Metropolitan Statistical Areas
Displayed difference is the percent RSE of the conventionally implemented linearization-based variance estimate minus the percent RSE of the corrected variance estimate.

Table 4
Mean percent change in estimates due to subnational recalibration by nationally weighted RSE and subnational area type

	States		MSAs	
Nationally Weighted Percent RSE	Num. Estimates	Mean Percent Change	Num. Estimates	Mean Percent Change
≤ 20%	438	3.5	638	4.7
20%-30%	156	5.1	434	6.3
30%-40%	69	6.7	337	8.7
40%-50%	25	12.3	188	11.0
> 50%	22	15.5	137	15.0
Overall	710	4.8	1734	7.4

Note: RSE = Relative Standard Error; MSAs=Metropolitan Statistical Areas
Percent change in estimate calculated as difference between recalibrated and nationally weighted point estimates divided by the nationally weighted estimate.

estimation methods, we examined the characteristics of estimates that have the largest differences in RSEs resulting from use of the corrected method over the conventionally implemented linearization-based method. That is, we identified when recalibration leads to the most measured efficiency gains for variance estimates. As shown in Table 4, recalibration resulted in a larger change for point estimates that were imprecise prior to recalibration compared to estimates that were already precise. For both states and MSAs, the mean percent change in point estimates increased as the nationally weighted (pre-subnational calibration) RSE increased.

Recalibration had a larger impact on the point estimates of imprecise estimates than precise estimates. This translated into larger precision gains resulting from use of the corrected calibration estimation method. Table 5 presents the mean change in the percent RSE for both the conventionally implemented linearization-based and corrected calibration estimation methods by the percent change in the point estimate resulting from subnational calibration. For both states and MSAs, estimates with a larger percent change due to recalibration had larger reductions in the percent RSE based on the corrected variance estimator compared to the original estimates. This trend was not observed with the conventionally implemented linearization-based variance estimator, as there appears to be no relationship between the change in the point estimate and the change in the percent RSE resulting from calibration.

Thus, the types of estimates where the recalibration provides the most precision gains are those that have lower precision to start with. This is a nice feature as these are the types of estimates that most need precision to be increased. With use of the conventionally implemented linearization-based estimator, many of these estimates would exhibit levels of precision requiring suppression or flagging as unreliable.

4 Conclusions and Recommendations

Calibration weighting is a powerful method for improving inference based on a sample when target population characteristics are known. The technique can be used to produce estimates for subdomains that a sample was not originally designed or weighted to produce (e. g., subnational estimates within a national survey) by controlling for differences between the weighting sample and population at the domain level. When the efficiencies provided by the calibration are appropriately accounted for in variance estimation, calibration can simultaneously reduce bias (or conditional bias) in point estimates while increasing the measured precision of these estimates. Conventionally implemented linearization-based variance estimation methods that apply calibrated weights to calculate survey estimates in a separate step following calibration are biased because they do not account for these efficiencies and tend to lead to artificially high variance estimates.

Replication-based variance estimators can be used to appropriately account for efficiencies in the weight calibration model. However, they require the running calibration-weighting routines in every replicate and the risk of failing to calibrate within some replicates. Because of these limitations, we conducted recalibration together with a corrected variance estimator to appropriately account for the efficiencies in our calibration models and calculate point and variance estimates at the time of calibration.

Whenever possible, researchers should consider using a version of calibration weighting that calculates linearization-based variance estimates at the time of calibration. Although the observed differences in precision estimates in our evaluation of subnational estimates were quite small, use of recalibration together with the corrected variance estimator resulted in variance estimates that tended to be smaller than those produced using the national calibration weights or recalibration with conventionally implemented linearization-based variance estimation. This led to fewer estimates flagged as unreliable based on estimated variance criteria.

Although the corrected variance estimator is better than the conventionally implemented linearization-based method, it has practical limitations. Analysts must have access to both the original (pre-calibration) weights as well as the control totals. There are circumstances when this is not possible. For example, public use datasets often include the final, calibrated survey weights and survey outcomes but do not provide the original weights or control totals. Fortunately in these cases, although conventionally implemented linearization-based variance estimates are biased, they tend to be biased upwards (artificially high), which increases the Type II error rate but not the Type I error rate.

It is also worth noting that while the increased efficiencies resulting from recalibration and measured with the corrected variance estimator were relatively small for the subnational NCVS data, results could vary when the methods are applied to other survey estimates. The efficiencies are a function of the original weights and the relationship between the survey value and the covariates in the calibration model.

Acknowledgements

The authors would like to thank the U.S. Bureau of Justice Statistics (BJS) for sponsoring this research. 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.

National Crime Victimization Survey (NCVS) data were analyzed and released from the Triangle Census Research Data Center. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed.

Table 5
Mean change in percent RSE for conventionally implemented linearization-based and corrected estimation methods by percent change in estimate and subnational area type

	States			MSAs		
Pct. Change in Estimate	Num. Estimates	Mean Change in Pct. RSE, Conventional	Mean Change in Pct. RSE, Corrected	Num. Estimates	Mean Change in Pct. RSE, Conventional	Mean Change in Pct. RSE, Corrected
≤ 2.5%	235	0.3	0.0	401	0.6	0.1
2.5% - 5.0%	238	0.2	0.0	403	0.6	0.0
5.0% - 7.5%	119	0.2	0.0	300	0.4	−0.2
7.5% - 10%	45	0.0	−0.5	234	0.4	−0.2
> 10%	73	−1.1	−2.2	396	0.8	−0.7
Overall	710	0.1	−0.2	1734	0.6	−0.2

Note: RSE = Relative Standard Error; MSAs = Metropolitan Statistical Areas

References

Deville, J.-C. & Sãrndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, *87*(418), 376–382.

Folsom, R. E. & Singh, A. C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. In *Proceedings of the american statistical association, survey research methods section* (Vol. 598603).

Holt, D. & Smith, T. F. (1979). Post stratification. *Journal of the Royal Statistical Society. Series A (General)*, 33–46.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, *32*(2), 133–142.

Little, R. & Vartivarian, S. (2006). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, *31*(2), 161–168.

Oh, H. & Scheuren, F. S. (1983). Weighting adjustments for unit nonresponse. In W. Madow, I. Olkin, & D. Rubin (Eds.), *Incomplete data in sample surveys* (Vol. 2, pp. 143–184). New York: Academic Press.

Research Triangle Institute. (2012). Sudaan language manual, volumes 1 and 2, release 11. Research Triangle Park, NC, Research Triangle Institute.

Royall, R. M. (1971). Linear regression models in finite population sampling theory. In D. Godambe V.P.and Sprott (Ed.), *Foundations of statistical inference* (pp. 259–279). Toronto: Holt, Rinehart, and Winston.

United States Department of Justice, Office of Justice Programs, & Bureau of Justice Statistics. (2010). National Crime Victimization Survey codebook, 2010 (ICPSR31202-v2). Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. Retrieved from http://doi.org/10.3886/ICPSR31202.v2

United States Department of Justice, Office of Justice Programs, & Bureau of Justice Statistics. (2014). Annual Survey of Jails: Jail-level data. [Computer file]. Conducted by U.S. Dept. of Commerce, Bureau of the Census. ICPSR36274-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producer and distributor].

Valliant, R., Dever, J. A., & Kreuter, F. (2013). *Practical tools for designing and weighting survey samples*. New York: Springer.

Ward, B., Clarke, T., Nugent, C., & Schiller, J. (2016). Early release of selected estimates based on data from the 2015 national health interview survey. National Center for Health Statistics. Retrieved from http://www.cdc.gov/nchs/nhis.htm

Williams, R. L., Heller, D., Couzens, G., Shook-Sa, B., Berzofsky, M., Smiley McDonald, H., & Krebs, C. (2015). Evaluation of direct variance estimation, estimate reliability, and confidence intervals for the national crime victimization survey. Prepared for the Bureau of Justice Statistics Research and Development Paper Series. Retrieved from https://www.bjs.gov/content/pub/pdf/edveercincvs.pdf