On Calibration to Estimated Totals in Survey Sampling

Takis Merkouris*

Abstract

Calibration of totals estimated from one survey to totals estimated from another survey is used in survey practice for consistency between these estimates and for reducing survey errors, e.g., undercoverage and nonresponse. We discuss issues of estimation efficiency of the resulting regression estimates for variables used in calibration and particularly for the rest of the survey variables, as well as practical problems with variance estimation. We point out that a calibration procedure that is statistically and operational more efficient is possible when micro-data from the other survey are available. In this procedure the combined survey data are calibrated simultaneously, so that estimated totals for common variables in the two surveys are calibrated to each other. We show that the improved efficiency of the regression estimates generated by this calibration procedure is due to the fact that the regression coefficients are approximately variance minimizing coefficients incorporating data from the two surveys. We also indicate that computations and variance estimation are greatly facilitated. An empirical study confirming the merits of the proposed calibration is also presented.

Key Words: Aligned estimates, composite calibration, optimal estimator, regression estimator, survey data combination, survey errors.

1. Introduction

Calibration in survey sampling is the part of the estimation process in which the sampling weights are adjusted to reproduce totals of auxiliary variables that are exactly known from registers or administrative sources. This procedure is extensively used to reduce non-sampling survey errors, e.g., undercoverage and nonresponse, and to improve the efficiency of estimators.

A variant of this standard calibration procedure in which the totals are estimates from other surveys is also used in current survey practice; examples of such calibration are given in Dever and Valliant (2017). Typically, this calibration setting involves two surveys, a primary survey and a "benchmark" survey, which have some common auxiliary variables with the same unknown totals. The primary objective of such calibration is consistency between estimates of totals from the two surveys for these auxiliary variables. Reduction of non-sampling survey errors may also be a motive if there is indication that for the specific variables such errors are more serious in the primary survey, but reduction of variance of the resulting estimators may not be realized for either the auxiliary variables or any target variable. The customary practice is to ignore the additional variability due to the estimated calibration totals when estimating the variance of estimators. However, valid variance estimation requires that the variance contribution due to estimating the calibration totals be accounted for. Recent research has proposed methods for incorporating variability of calibration totals into the variance estimation when the necessary information from the benchmark survey is available; see Dever and Valliant (2010), Opsomer and Erciulescu (2021), Guandalini and Tillé (2017). These methods, however, are rather complicated, and require specialized computer programs.

*Athens University of Business and Economics, 2 Trias Street, Athens 11362, Greece, merkouris@aueb.gr

In the existing literature there is lack of detailed study of the structure and properties of estimators generated from calibration using estimated totals. Issues such as the variance of estimators for both the auxiliary variables and target variables as functions of their correlation and the relative sample size of the primary and benchmark survey have not been investigated.

In this article we investigate the effect of calibration to estimated totals on the properties of the resulting estimators of totals for both the auxiliary and the target variables. We derive analytic results on the efficiency of these estimators in the case that standard calibration generates generalized regression estimator or optimal regression estimator. In situations where micro-data from the benchmark survey are available, then a calibration procedure that is statistically and operationally more efficient is possible. In this procedure the weights of the combined samples of the two surveys are calibrated simultaneously, so that estimated totals for common variables in the two surveys are calibrated to each other. Literature on relevant theory and practice for this type of calibration is available; see Merkouris (2004, 2013). The merits of this approach relative to the current procedure are studied both theoretically and through a simulation study.

2. Calibration and regression estimation

2.1 Calibration to known totals

Let $U = \{1, \cdots, i, \cdots, N\}$ denote a finite population of N units, and let s denote a sample of size n drawn from the population U, using a sampling design that defines inclusion probability $\pi_i = P(i \in s)$ for unit $i \in U$, and joint inclusion probability $\pi_{ij} = P(i, j \in s)$ for units $i, j \in U$. Assuming that $\pi_i > 0$ for all $i \in U$, the design weight of unit $i \in s$ is $w_i = 1/\pi_i$. For any variable of interest y, with values y_i, $i \in U$, and population total $Y = \sum_U y_i$, the Horvitz-Thompson (HT) estimator of Y is defined as $\hat{Y} = \sum_s w_i y_k$.

Consider first the standard calibration procedure whereby the weights w_i are adjusted to reproduce known totals of a number of auxiliary variables. To facilitate theoretical and empirical analysis we consider here the case of calibration involving a single auxiliary variable x with known total $X = \sum_U x_i$, and HT estimator $\hat{X} = \sum_s w_i x_i$. A calibration procedure transforms the weights w_i to weights c_i that satisfy the constrain $\sum_s c_i x_i = X$ while minimizing a distance function between the weights w_i and c_i. Calibration using the distance function $\sum_s (c_i - w_i)^2/w_i$ produces calibrated weights given in closed form by $c_i = w_i + w_i x_i (\sum_s w_i x_i^2)^{-1}(X - \hat{X})$. These calibrated weights generate the linear estimator $\hat{Y}^R = \sum_s c_i y_i$ of Y, given also in the form of generalized regression estimator as

$$\hat{Y}^R = \hat{Y} + \hat{\beta}(X - \hat{X}), \tag{1}$$

where $\hat{\beta} = \sum_s w_i y_i x_i / \sum_s w_i x_i^2$. This most commonly used calibration procedure, leading to the regression estimator $\hat{Y}^R$, will help to derive exact results in the rest of this article. For extensive discussion on the generalized regression estimation see Deville and Särndal (1992), Särndal et al. (1992), Montanari (1998), Fuller (2002), Särndal (2007).

The estimator $\hat{Y}^R$ is asymptotically (i.e., for large samples) unbiased, i.e., $\mathrm{E}(\hat{Y}^R) = Y$, and its asymptotic variance of $\hat{Y}^R$ is given by

$$\mathrm{Var}(\hat{Y}^R) = \mathrm{Var}(\hat{Y}) + \beta^2 \mathrm{Var}(\hat{X}) - 2\beta \mathrm{Cov}(\hat{Y}, \hat{X}),$$

where $\beta = \sum_U y_i x_i / \sum_U x_i^2$ is the population version of $\hat{\beta}$.

By the calibration property, $\hat{X}^R = X$, and thus $\mathrm{Var}(\hat{X}^R) = 0$. Besides ensuring consistency between $\hat{X}^R$ and X, calibration may increase the estimation efficiency, yielding $\mathrm{Var}(\hat{Y}^R) < \mathrm{Var}(\hat{Y})$, if y and x are highly correlated.

2.2 Calibration to estimated totals

Now suppose that the total X is not known, and $\hat{X}$ is calibrated to an estimate $\tilde{X}$ of X from another survey, the "benchmark" survey, with sample $s^\star$; it is assumed that the expected value of both $\hat{X}$ and $\tilde{X}$ is X. This calibration, which treats $\tilde{X}$ as a constant calibration total, generates the regression-type estimator of Y

$$\tilde{Y}^R = \hat{Y} + \hat{\beta}(\tilde{X} - \hat{X}), \tag{2}$$

where $\hat{\beta}$ remains the same as in (1).

Again, by calibration $\tilde{X}^R = \tilde{X}$, and thus $\text{Var}(\tilde{X}^R) = \text{Var}(\tilde{X})$. Obviously, the calibration estimator $\tilde{X}^R$ will have larger variance than the basic estimator $\hat{X}$ if $\text{Var}(\tilde{X}) > \text{Var}(\hat{X})$.

The asymptotic variance of $\tilde{Y}^R$ is

$$\begin{aligned} \text{Var}(\tilde{Y}^R) &= \text{Var}(\hat{Y}) + \beta^2\text{Var}(\hat{X}) - 2\beta\text{Cov}(\hat{Y}, \hat{X}) + \beta^2\text{Var}(\tilde{X}) \\ &= \text{Var}(\hat{Y}^R) + \beta^2\text{Var}(\tilde{X}). \end{aligned}$$

It is obvious that using the estimate $\tilde{X}$ as calibration total increases the variance of the resulting estimator $\tilde{Y}^R$ by the quantity $\beta^2\text{Var}(\tilde{X})$, relative to the standard regression estimator which involves the constant X as calibration total. It may well happen that the estimator $\tilde{Y}^R$ has larger variance than the basic estimator $\hat{Y}$ even with high correlation of y with x, and this variance is not estimated correctly if an estimate $\widehat{\text{Var}}(\tilde{X})$ of the variance $\text{Var}(\tilde{X})$ is not available (along with $\tilde{X}$). It is customary in survey practice to ignore the variability of $\tilde{X}$ when estimating $\text{Var}(\tilde{Y}^R)$. In case that accounting for this additional variability is possible, valid estimation of $\text{Var}(\tilde{Y}^R)$ is possible, though complicated; see Dever and Valliant (2017) and Opsomer (2021).

It should be noted that the estimator $\tilde{Y}^R$ is not a proper regression estimator because, unlike the regression estimator $\hat{Y}^R$ in (1), the regression coefficient $\hat{\beta}$ is not a function of all survey data in the zero-expectation term $\tilde{X} - \hat{X}$. Furthermore, the condition that $\text{E}(\tilde{X}) = \text{E}(\hat{X})$ is crucial for $\tilde{Y}^R$ to be asymptotically unbiased. Henceforth the estimator $\tilde{Y}^R$ will be referred to as pseudo-regression estimator.

The effect of using the estimate $\tilde{X}$ as calibration total is analytically assessed in the case of the optimal regression estimation, which though is more complicated than the generalized regression estimation, and practicable only for certain sampling designs (i.e., simple random sampling, Poisson sampling and their stratified versions). The optimal regression estimator (henceforth referred to as optimal estimator) is given by

$$\hat{Y}^O = \hat{Y} + \hat{\beta}^O(X - \hat{X}), \tag{3}$$

where $\hat{\beta}^O = \widehat{\text{Cov}}(\hat{Y}, \hat{X})/\widehat{\text{Var}}(\hat{X})$ is the optimal regression coefficient, inducing a regression estimator $\hat{Y}^O = \hat{Y} + \hat{\beta}^O(X - \hat{X})$ with minimum asymptotic variance; see, for example, Montanari (1987, 1998) and Rao (1994). The coefficient $\hat{\beta}^O$ is an estimate of the population regression coefficient $\beta^O = \text{Cov}(\hat{Y}, \hat{X})/\text{Var}(\hat{X})$, and its explicit expression is $\hat{\beta}^O = \sum_s \sum_s v_{ij} y_i x_j (\sum_s \sum_s v_{ij} x_i x_j)^{-1}$, where $v_{ij} = (\pi_{ij} - \pi_i\pi_j)/\pi_{ij}\pi_i\pi_j$; $\pi_{ii} = \pi_i$. The estimator $\hat{Y}^O$ is generated by a calibration procedure that minimizes a special distance function between design and calibrated weights; see, for example, Andersson and Thorburn (2005). It can be written in linear form as $\hat{Y}^O = \sum_s c_i^O y_i$, with the calibrated weights c_i^O given by $c_i^O = w_i + \sum_{s_{j\neq i}} v_{ij} x_i (\sum_s \sum_s v_{ij} x_i x_j)^{-1}(X - \hat{X})$.

The asymptotic variance of $\hat{Y}^O$ is given by

$$\begin{aligned} \text{Var}(\hat{Y}^O) &= \text{Var}(\hat{Y}) + {\beta^O}^2\text{Var}(\hat{X}) - 2\beta^O\text{Cov}(\hat{Y}, \hat{X}) \\ &= \text{Var}(\hat{Y}) - \frac{\text{Cov}^2(\hat{Y}, \hat{X})}{\text{Var}(\hat{X})}. \end{aligned}$$

In analogy to the estimator $\tilde{Y}^R$, if we replace the calibration total X by the estimate $\tilde{X}$ in (3), we obtain the estimator

$$\tilde{Y}^O = \hat{Y} + \hat{\beta}^O(\tilde{X} - \hat{X}), \tag{4}$$

where $\hat{\beta^O}$ is the same as in (3). The asymptotic variance of $\tilde{Y}^O$ is

$$\begin{aligned} \text{Var}(\tilde{Y}^O) &= \text{Var}(\hat{Y}^O) + {\beta^O}^2\text{Var}(\tilde{X}) \\ &= \text{Var}(\hat{Y}) - \frac{\text{Cov}^2(\hat{Y}, \hat{X})}{\text{Var}(\hat{X})}\left[1 - \frac{\text{Var}(\tilde{X})}{\text{Var}(\hat{X})}\right]. \end{aligned} \tag{5}$$

Clearly, $\text{Var}(\tilde{Y}^O) > \text{Var}(\hat{Y}^O)$. Note, though that $\text{Var}(\tilde{Y}^O) = \text{Var}(\hat{Y})$ if $\text{Var}(\tilde{X}) = \text{Var}(\hat{X})$, while $\text{Var}(\tilde{Y}^O) < \text{Var}(\hat{Y})$ only if $\text{Var}(\tilde{X}) < \text{Var}(\hat{X})$. Rewriting (5), in terms of the efficiency of $\tilde{Y}^O$ relative to $\hat{Y}$, as

$$\frac{\text{Var}(\tilde{Y}^O) - \text{Var}(\hat{Y})}{\text{Var}(\hat{Y})} = -\rho^2(\hat{Y}, \hat{X})\left[1 - \frac{\text{Var}(\tilde{X})}{\text{Var}(\hat{X})}\right],$$

we see that when $\text{Var}(\tilde{X}) > \text{Var}(\hat{X})$ we get

$$\frac{\text{Var}(\tilde{Y}^O) - \text{Var}(\hat{Y})}{\text{Var}(\hat{Y})} > 0,$$

that is, $\tilde{Y}^O$ is less efficient than the basic estimator $\hat{Y}$. The larger than one is the ratio $\text{Var}(\tilde{X})/\text{Var}(\hat{X})$, and the larger the correlation $\rho(\hat{Y}, \hat{X})$, the larger the inefficiency of $\tilde{Y}^O$. In simple random sampling without replacement we get $\rho(\hat{Y}, \hat{X}) = \rho(y, x)$, i.e., the estimator $\tilde{Y}^O$ loses efficiency as the correlation of y with x increases. This absurdity is likely to hold in other sampling schemes, where the property $\rho(\hat{Y}, \hat{X}) = \rho(y, x)$ does not hold exactly.

By the calibration property, $\tilde{X}^O = \tilde{X}$, and thus $\text{Var}(\tilde{X}^O) = \text{Var}(\tilde{X})$. Hence, the optimal regression estimator $\tilde{X}^O$ will have larger variance than the basic estimator $\hat{X}$ if $\text{Var}(\tilde{X}) > \text{Var}(\hat{X})$.

3. Composite calibration and regression estimation

The estimator (4) is in fact a pseudo-optimal estimator, because it is constructed as a calibration estimator with an estimated total as calibration total while retaining the regression coefficient $\hat{\beta}^O$. The optimal regression coefficient for the estimator in (4) is $\hat{\beta}^{CO} = \widehat{\text{Cov}}(\hat{Y}, \hat{X})/\left[\widehat{\text{Var}}(\hat{X}) + \widehat{\text{Var}}(\tilde{X})\right]$. This coefficient incorporates the data from the benchmark survey, and the variability of the estimate $\tilde{X}$, giving the composite optimal regression estimator

$$\tilde{Y}^{CO} = \hat{Y} + \hat{\beta}^{CO}(\tilde{X} - \hat{X}), \tag{6}$$

which combines information from the two surveys. The asymptotic variance of $\tilde{Y}^{CO}$ is

$$\text{Var}(\tilde{Y}^{CO}) = \text{Var}(\hat{Y}) - \frac{\text{Cov}^2(\hat{Y}, \hat{X})}{\left[\text{Var}(\hat{X}) + \text{Var}(\tilde{X})\right]}.$$

Straightforward algebra shows that

$$\text{Var}(\tilde{Y}^{CO}) - \text{Var}(\tilde{Y}^O) = -\frac{\text{Cov}^2(\hat{Y}, \hat{X})}{\text{Var}^2(\hat{X})}\left[\frac{\text{Var}^2(\tilde{X})}{\text{Var}(\hat{X}) + \text{Var}(\tilde{X})}\right]. \tag{7}$$

Therefore, $\text{Var}(\tilde{Y}^{CO}) < \text{Var}(\tilde{Y}^{O})$ regardless of the relative sizes of the variances $\text{Var}(\hat{X})$ and $\text{Var}(\tilde{X})$. The expression in the right hand of (7) represents the reduction in variance when using $\text{Var}(\tilde{Y}^{CO})$ instead of $\text{Var}(\tilde{Y}^{O})$.

It follows that the composite optimal regression estimator of X is

$$\tilde{X}^{CO} = \hat{X} + \hat{\beta}_x^{CO}(\tilde{X} - \hat{X}) = \hat{\beta}_x^{CO}\tilde{X} + (1 - \hat{\beta}_x^{CO})\hat{X}, \tag{8}$$

where $\hat{\beta}_x^{CO} = \widehat{\text{Var}}(\hat{X}) / \left[\widehat{\text{Var}}(\hat{X}) + \widehat{\text{Var}}(\tilde{X})\right]$, i.e., the estimator $\tilde{X}^{CO}$ is a weighted average of the estimators $\tilde{X}$ and $\hat{X}$. In fact, $\tilde{X}^{CO}$ is asymptotically the best linear unbiased combination of $\tilde{X}$ and $\hat{X}$. The asymptotic variance of $\hat{\beta}_x^{CO}$ is

$$\text{Var}(\tilde{X}^{CO}) = \frac{\text{Var}(\hat{X})\text{Var}(\tilde{X})}{\left[\text{Var}(\hat{X}) + \text{Var}(\tilde{X})\right]}.$$

Clearly then $\text{Var}(\tilde{X}^{CO}) < \text{Var}(\tilde{X}^{O})$, with the reduction of the variance $\text{Var}(\tilde{X}^{CO}) - \text{Var}(\tilde{X}^{O})$ being equal to $\text{Var}^2(\tilde{X}) / \left[\text{Var}(\hat{X}) + \text{Var}(\tilde{X})\right]$. In percent, this amounts to a reduction by $\text{Var}(\tilde{X}) / \left[\text{Var}(\hat{X}) + \text{Var}(\tilde{X})\right]$.

The more efficient composite estimators (6) and (8) are generated through a proper calibration procedure involving the combined samples of the two surveys, whereby the estimates of X from the two surveys are calibrated to each other; see Merkouris (2013). This motivates the construction of more practical composite regression estimators of Y and X through a similar calibration procedure; see Merkouris (2004, 2013). The composite regression estimator of Y is given then by

$$\tilde{Y}^{CR} = \hat{Y} + \hat{\beta}^{CR}(\tilde{X} - \hat{X}), \tag{9}$$

where $\hat{\beta}^{CR} = \sum_s w_i y_i x_i / \left[\sum_s w_i x_i^2 + \sum_{s^\star} w_i x_i^2\right]$. Similarly, the composite regression estimator of X is given by

$$\tilde{X}^{CR} = \hat{X} + \hat{\beta}_x^{CR}(\tilde{X} - \hat{X}) = \hat{\beta}_x^{CR}\tilde{X} + (1 - \hat{\beta}_x^{CR})\hat{X}, \tag{10}$$

where $\hat{\beta}_x^{CR} = \sum_s w_i x_i^2 / \left[\sum_s w_i x_i^2 + \sum_{s^\star} w_i x_i^2\right]$.

It is to be noted that the regression coefficients $\hat{\beta}^{CR}$ and $\hat{\beta}_x^{CR}$ incorporate data from the two surveys, thus enhancing the efficiency of the regression estimators $\tilde{Y}^{CR}$ and $\tilde{X}^{CR}$. For the same reason these estimators are proper regression estimators, in contrast to the estimator $\tilde{Y}^{R}$ in (2) whose regression coefficient $\hat{\beta}$ is a function of data from the primary survey only. In effect, in the calibration of the combined samples that generates the estimator $\tilde{Y}^{CR}$, the difference $\tilde{X} - \hat{X}$, involving data from the two surveys, is calibrated to zero; see Merkouris (2004, 2013). It is instructive to rewrite (9) in the equivalent form

$$\tilde{Y}^{CR} = \hat{Y} + \hat{\beta}(\tilde{X}^{CR} - \hat{X}), \tag{11}$$

where $\hat{\beta}$ is as in (2). This would be the regression estimator generated by a calibration procedure based only on the sample s, but using as calibration total the composite estimator $\tilde{X}^{CR}$ instead of the single-survey estimator $\tilde{X}$ used as calibration total in the estimator $\tilde{X}^{R}$ in (2). However, the calibration of the combined samples of the two surveys is much more practical in generating $\tilde{Y}^{CR}$ and in estimating its variance.

The form of the weighting coefficient $\hat{\beta}_x^{CR}$ in the weighted average $\tilde{X}^{CR}$ in (10), as a approximation of the optimal coefficient $\hat{\beta}_x^{OR}$ in(8), suggests that $\tilde{X}^{CR}$ will have smaller variance than the regression estimator $\tilde{X}^{R} = \tilde{X}$. This can be proved exactly for $\text{Var}(\hat{X}) < \text{Var}(\tilde{X})$, and approximately (assuming that $\hat{\beta}_x^{CR} \approx n_s/(n_s + n_{s^\star})$) otherwise. The form (11) of $\tilde{Y}^{CR}$ suggests that this composite regression estimator has smaller variance than the regression estimator $\tilde{Y}^{R}$, but exact proof of this is not workable. An empirical evaluation of the efficiency of $\tilde{X}^{CR}$ and $\tilde{Y}^{CR}$ is presented in the next section.

4. Simulation Study

In this simulation study, we compared estimators of X and Y generated by calibration procedures that use estimated calibration totals. A vector (y, x) was specified to have the bivariate lognormal distribution with means $\mathrm{E}(y) = 8$, $\mathrm{E}(x) = 5$ and pairs of variances $\mathrm{Var}(y) = (10, 50)$, $\mathrm{Var}(x) = (10, 50)$, with associated coefficients of variation $\mathrm{CV}(y) = (0.40, 0.88)$, $\mathrm{CV}(x) = (0.63, 1.41)$. For each of these four parameter combinations we specified three correlations for (y, x), namely $\rho(y, x) = (0.25, 0.50, 0.75)$, thus defining 12 different distributions for (y, x) to test the effect of using an estimated total as calibration total on the efficiency of the OR and GR estimators for such varying distributional settings. For each of these 12 distributions, a population of size $N = 1000000$ was simulated by generating values of the vector (y, x). In addition, to reflect the relative sizes of $\mathrm{Var}(\tilde{X})$ and $\mathrm{Var}(\hat{X})$, three combinations of sizes $(n_{s^\star}, n_s)$ of two samples $s^\star$ and s were specified, namely, (500, 5000), (5000, 5000) and (5000, 500), thus creating a total of 36 simulation settings.

Two different sampling designs were considered for selecting, independently, the samples s and $s^\star$, namely simple random sampling (SRS) without replacement and Poisson sampling. In both SRS and Poisson sampling the optimal estimator can be computed. Also, SRS is an equal probability (of selection) sampling design, whereas Poisson is an unequal probability sampling design, with x being used in defining the selection probabilities as well as in estimation.

First, using SRS we drew $r = 20000$ pairs of samples $(s^\star, s)$ from each of the 36 simulated populations, and with each drawing we generated the estimates $\hat{X}$, $\tilde{X}$, $\tilde{X}^{CR}$, $\tilde{X}^{CO}$, $\hat{Y}$, $\tilde{Y}^{R}$, $\tilde{Y}^{CR}$, $\tilde{Y}^{O}$, $\tilde{Y}^{CO}$ (as defined in Sections 2 and 3). Using the r replicates of these estimates we calculated the empirical variances of all these estimators and the relative bias of the estimators $\tilde{X}^{CR}$, $\tilde{X}^{CO}$, $\tilde{Y}^{R}$, $\tilde{Y}^{CR}$, $\tilde{Y}^{O}$, $\tilde{Y}^{CO}$. The efficiency of each of the estimators $\tilde{X}^{CR}$ and $\tilde{X}^{CO}$ is assessed through the relative difference between its variance and the variance of the estimator $\tilde{X}$, and the efficiency of each of the estimators $\tilde{Y}^{R}$, $\tilde{Y}^{CR}$, $\tilde{Y}^{O}$, $\tilde{Y}^{CO}$ is assessed through the relative difference between its variance and the variance of the estimator $\hat{Y}$. For example, for $\tilde{Y}^{R}$ the relative difference is $[\mathrm{Var}(\tilde{Y}^{R}) - \mathrm{Var}(\hat{Y})]/\mathrm{Var}(\hat{Y})$; this relative difference shows the reduction of the variance of $\tilde{Y}^{R}$ relative to the variance of the basic estimator $\hat{Y}$.

In the following tables, the efficiencies (%) of the estimators $\tilde{X}^{CR}$, $\tilde{X}^{CO}$ relative to the estimator $\tilde{X}$ are displayed under the headings $\tilde{X}^{CR}|\tilde{X}$ and $\tilde{X}^{CO}|\tilde{X}$, respectively, and the efficiencies of the estimators $\tilde{Y}^{R}$, $\tilde{Y}^{CR}$, $\tilde{Y}^{O}$, $\tilde{Y}^{CO}$ relative to the estimator $\hat{Y}$ are displayed under the headings $\tilde{Y}^{R}|\hat{Y}$, $\tilde{Y}^{CR}|\hat{Y}$, $\tilde{Y}^{O}|\hat{Y}$, and $\tilde{Y}^{CO}|\hat{Y}$, respectively. Negative sign indicates efficiency gain (percent reduction of variance).

As shown in Table 1, the efficiency of $\tilde{X}^{CR}$ is around 91%, 50% and 9.5% for the sample size settings (500, 5000), (5000,5000) and (5000, 500), respectively, in all four distributions and irrespective of the three correlations $\rho(y, x)$. This performance of the composite regression estimator $\tilde{X}^{CR}$ is explained by its construction as a weighted combination of the estimators $\hat{X}$ and $\tilde{X}$, with weights reflecting the relative sample sizes of the two samples.

Regarding the efficiency of the estimator $\tilde{Y}^{R}$, for the sample size setting (500, 5000) this estimator is vastly less efficient than the basic estimator $\hat{Y}$, more so as the cv and skewness of x gets larger than the cv and skewness of y, and as the correlation $\rho(y, x)$ increases. The estimator $\tilde{Y}^{R}$ shows some bias (not reported in Table 1), ranging from a negative maximum of -4.0% (relative to Y) in distribution 3 to a positive maximum of 4.7% in distribution 1.

For the sample size setting (5000, 5000), when $Var(\tilde{X}) \approx Var\hat{X})$, the estimator $\tilde{Y}^{R}$ is less inefficient than the estimator $\hat{Y}$, but not as severely as before. This inefficiency is as

before with respect to the relative size of cv of x and y, but decreases as $\rho(y,x)$ increases. The estimator $\tilde{Y}^R$ is more efficient than $\hat{Y}$, by 2.3%, only in one of the twelve simulation settings (4 distributions by 3 correlations). The bias of $\tilde{Y}^R$ is smaller then before, with a maximum of 1.5%.

As expected, for the third sample size setting (5000, 500), when $Var(\tilde{X}) < Var\hat{X})$, the inefficiency of $\tilde{Y}^R$ is even less severe. Its pattern with respect to the cv of y and x and the correlation $\rho(y,x)$ (except for sharper decrease as $\rho(y,x)$ increases) is as in the setting (5000, 5000), but now $\tilde{Y}^R$ is more efficient than $\hat{Y}$ in three simulation settings (in distributions 1 and 2), by a maximum of 51%. The bias of $\tilde{Y}^R$ ranges from a minimum of -0.41% to a maximum of 2%.

The efficiency of the composite regression estimator $\tilde{Y}^{CR}$, relative to $\hat{Y}$, increases as $\rho(y,x)$ increases, more sharply as we move from (500, 5000) to (5000, 500). The estimator $\tilde{Y}^{CR}$ is more efficient than $\hat{Y}$ in five of the twelve simulation settings in each of the three sample size settings, mostly for higher cv of y. These efficiency gains get larger as we move from (500, 5000) to (5000, 500), with highest gain of 5% in (500, 5000), 28% in (5000, 5000) and 51% in (5000, 500). The empirical bias of $\tilde{Y}^{CR}$ is negligible, comparable to the empirical bias of $\tilde{Y}^{CR}$. Finally, $\tilde{Y}^{CR}$ is more efficient than the pseudo regression estimator $\tilde{Y}^R$ in all settings, more vastly as $n_{s_\star}$ gets smaller than n_s, and as we move from distribution 1 to distribution 4. This efficiency is not reported in Table 1, but can be easily derived from the reported efficiencies $\tilde{Y}^R|\hat{Y}$ and $\tilde{Y}^{CR}|\hat{Y}$.

Table 2 displays the efficiencies of $\tilde{X}^{CO}$, $\tilde{Y}^O$ and $\tilde{Y}^{CO}$. The efficiency of $\tilde{X}^{CO}$ is almost identical to that of $\tilde{X}^{CR}$ (displayed in Table 1) in all 36 simulation settings.

For the sample size setting (500, 5000), the estimator $\tilde{Y}^O$ is very inefficient relative to the basic estimator $\hat{Y}$, but much less than the regression estimator $\tilde{Y}^R$; the performance of $\tilde{Y}^O$ is very similar across the four distributions. A clear loss of efficiency as the correlation $\rho(y,x)$ increases is shown, confirming the oddity noted in Section 2.2. For the sample size setting (5000, 5000) the empirical variances of $\tilde{Y}^O$ and $\hat{Y}$ are about the same, confirming the theoretical property $\text{Var}(\tilde{Y}^O) = \text{Var}(\hat{Y})$ shown in Section 2.2. For the setting (5000, 500), $\tilde{Y}^O$ is more efficient than $\hat{Y}$ in all 12 other simulation settings, the efficiency increasing with increasing correlation $\rho(y,x)$ but being stable across the four distributions.

The composite optimal estimator $\tilde{Y}^{CO}$ is more efficient than the basic estimator $\hat{Y}$ in all 36 simulation settings, increasingly so as we move from the samples (500, 5000) to the samples (5000, 500) and from low to high correlation $\rho(y,x)$. The performance of $\tilde{Y}^{CO}$ is very similar across the four distributions. Furthermore, $\tilde{Y}^{CO}$ is more efficient than $\tilde{Y}^{CR}$ in all 36 simulation settings, and more efficient than $\tilde{Y}^O$ in all settings, more so as $n_{s^\star}$ gets larger than n_s.

The results for the Poisson sampling are displayed in Tables 3 and 4.

As shown in Table 3, the efficiency of $\tilde{X}^{CR}$ is very similar to that in SRS, only very slightly weaker for the samples (5000, 500).

The estimator $\tilde{Y}^R$ is quite inefficient relative to $\hat{Y}$ for the samples (500, 5000), but less than in SRS. The inefficiency of $\tilde{Y}^R$ increases as $\rho(y,x)$ increases and decreases as we move from distribution 1 to distribution 4. For the samples (5000, 5000) and (5000, 500) $\tilde{Y}^R$ is more efficient than $\hat{Y}$ (as expected for the inefficient Poisson sampling) in all simulation settings. The efficiency of $\tilde{Y}^R$ is higher for the samples (5000, 500), more so for increasing $\rho(y,x)$ and in distributions 1 and 2. The bias of $\tilde{Y}^R$ reaches a negative maximum of -4.3%.

The composite regression estimator $\tilde{Y}^{CR}$ is more efficient than $\hat{Y}$ in all 36 simulation settings, more in distributions 1 and 2. The efficiency increases as $n_{s^\star}$ gets larger than n_s and as $\rho(y,x)$ increases. The estimator $\tilde{Y}^{CR}$ is more efficient than $\tilde{Y}^R$, more as $n_{s_\star}$ gets smaller than n_s, except for the samples (5000, 500) in distributions 1 and 2 where the two

Table 1: Efficiency (%) of the $\tilde{X}^{CR}$, $\tilde{Y}^{R}$ and $\tilde{Y}^{CR}$ estimators (SRS)

$\rho(\mathbf{y}.\mathbf{x})$	$(n_{s^\star} = 500.n_s = 5000)$			$(n_{s^\star} = 5000.n_s = 5000)$			$(n_{s^\star} = 5000.n_s = 500)$		
	$\tilde{\mathbf{X}}^{\mathbf{CR}}\|\tilde{\mathbf{X}}$	$\tilde{\mathbf{Y}}^{\mathbf{R}}\|\hat{\mathbf{Y}}$	$\tilde{\mathbf{Y}}^{\mathbf{CR}}\|\hat{\mathbf{Y}}$	$\tilde{\mathbf{X}}^{\mathbf{CR}}\|\tilde{\mathbf{X}}$	$\tilde{\mathbf{Y}}^{\mathbf{R}}\|\hat{\mathbf{Y}}$	$\tilde{\mathbf{Y}}^{\mathbf{CR}}\|\hat{\mathbf{Y}}$	$\tilde{\mathbf{X}}^{\mathbf{CR}}\|\tilde{\mathbf{X}}$	$\tilde{\mathbf{Y}}^{\mathbf{R}}\|\hat{\mathbf{Y}}$	$\tilde{\mathbf{Y}}^{\mathbf{CR}}\|\hat{\mathbf{Y}}$
				Distribution 1: CV(y)=0.88. CV(x)=0.63					
0.25	-90.84	343.84	0.47	-50.18	38.99	2.53	-9.76	8.90	4.87
0.50	-90.93	408.05	-2.05	-50.39	20.82	-11.03	-9.71	-17.83	-20.19
0.75	-91.03	479.66	-5.06	-50.35	-2.30	-27.53	-9.58	-50.67	-50.90
				Distribution 2: CV(y)=0.40. CV(x)=0.63					
0.25	-91.13	1.586.71	8.51	-50.02	228.20	41.68	-9.44	104.41	80.93
0.50	-91.13	1.721.17	4.17	-50.07	199.47	18.23	-9.48	55.38	34.84
0.75	-91.13	1.861.56	-0.78	-50.03	164.85	-8.88	-9.45	0.45	-16.70
				Distribution 3: CV(y)=0.88. CV(x)=1.41					
0.25	-91.12	529.09	1.90	-49.76	62.63	6.42	-9.84	24.87	15.90
0.50	-91.12	775.11	-0.22	-49.78	64.49	-5.78	-9.85	0.52	-8.61
0.75	-91.11	1062.49	-3.39	-49.64	59.78	-23.79	-9.77	-34.59	-43.10
				Distribution 4: CV(y)=0.40. CV(x)=1.41					
0.25	-91.09	2.044.43	13.67	-49.36	299.50	55.73	-9.54	161.37	120.27
0.50	-91.09	2.506.47	10.87	-49.39	313.57	38.59	-9.57	127.34	84.54
0.75	-91.09	3.017.96	7.05	-49.37	324.02	16.08	-9.55	81.56	37.99

estimators have almost identical efficiency.

Table 4 displays the efficiencies of $\tilde{X}^{CO}$, $\tilde{Y}^{O}$ and $\tilde{Y}^{CO}$. The efficiency of $\tilde{X}^{CO}$ is almost identical to that of $\tilde{X}^{CR}$ (displayed in Table 3) in all 36 simulation settings.

For the samples (500, 5000), the optimal estimator $\tilde{Y}^{O}$ is very inefficient relative to $\hat{Y}$, more so as $\rho(y, x)$ increases. For the samples (5000, 5000) the estimator $\tilde{Y}^{O}$ is very slightly inefficient. For the samples (5000, 500), $\tilde{Y}^{O}$ is more efficient than $\hat{Y}$ in all 12 other simulation settings (more in distributions 1 and 2), the efficiency increasing with increasing correlation $\rho(y, x)$. For the samples (5000, 500) the efficiency of $\tilde{Y}^{O}$ is similar to that of $\tilde{Y}^{R}$, but a little higher in distributions 3 an 4. It is also much higher than in the case of SRS.

The composite optimal estimator $\tilde{Y}^{CO}$ is more efficient than the basic estimator $\hat{Y}$ in all 36 simulation settings, increasingly so as we move from the samples (500, 5000) to the samples (5000, 500) and from low to high correlation $\rho(y, x)$. In the samples (5000, 5000) and (5000, 500) the efficiency of $\tilde{Y}^{CO}$ is higher in distributions 1 and 2. The efficiency of $\tilde{Y}^{CO}$ is comparable to that of $\tilde{Y}^{CR}$ in most setting; it is a litle higher in samples (5000, 5000), (5000, 500) and in distributions 2 and 3. Finally, $\tilde{Y}^{CO}$ is considerably more efficient than $\tilde{Y}^{O}$ in samples (500, 5000), (5000, 5000), but almost as efficient in samples (5000, 500).

Table 2: Efficiency (%) of the $\tilde{X}^{CO}$, $\tilde{Y}^{O}$ and $\tilde{Y}^{CO}$ estimators (SRS)

$\rho(\mathbf{y}.\mathbf{x})$	$(n_{s^\star}=500.n_s=5000)$			$(n_{s^\star}=5000.n_s=5000)$			$(n_{s^\star}=5000.n_s=500)$		
	$\mathbf{\tilde{X}^{CO}}\lvert\mathbf{\tilde{X}}$	$\mathbf{\tilde{Y}^{O}}\lvert\mathbf{\hat{Y}}$	$\mathbf{\tilde{Y}^{CO}}\lvert\mathbf{\hat{Y}}$	$\mathbf{\tilde{X}^{CO}}\lvert\mathbf{\tilde{X}}$	$\mathbf{\tilde{Y}^{O}}\lvert\mathbf{\hat{Y}}$	$\mathbf{\tilde{Y}^{CO}}\lvert\mathbf{\hat{Y}}$	$\mathbf{\tilde{X}^{CO}}\lvert\mathbf{\tilde{X}}$	$\mathbf{\tilde{Y}^{O}}\lvert\mathbf{\hat{Y}}$	$\mathbf{\tilde{Y}^{CO}}\lvert\mathbf{\hat{Y}}$
				Distribution 1: CV(y)=0.88. CV(x)=0.63					
0.25	-90.83	56.58	-0.58	-50.16	0.15	-3.07	-9.66	-5.10	-5.29
0.50	-90.91	227.83	-2.28	-50.36	0.61	-12.29	-9.60	-21.75	-22.18
0.75	-91.02	518.41	-5.01	-50.31	1.50	-27.52	-9.47	-50.03	-50.70
				Distribution 2: CV(y)=0.40. CV(x)=0.63					
0.25	-91.12	57.77	-0.50	-49.99	-0.43	-3.27	-9.34	-4.99	-5.17
0.50	-91.12	230.53	-2.04	-50.04	-0.02	-12.30	-9.38	-21.68	-22.10
0.75	-91.12	519.04	-4.64	-50.00	0.67	-27.50	-9.35	-49.85	-50.59
				Distribution 3: CV(y)=0.88. CV(x)=1.41					
0.25	-91.07	60.12	-0.57	-49.56	-0.09	-3.23	-9.40	-4.21	-4.79
0.50	-91.07	238.50	-2.06	-49.58	0.62	-12.37	-9.41	-20.58	-21.83
0.75	-91.06	531.37	-4.51	-49.43	1.09	-27.74	-9.32	-48.98	-50.72
				Distribution 4: CV(y)=0.40. CV(x)=1.41					
0.25	-91.04	60.25	-0.46	-49.15	-0.46	-3.39	-9.08	-4.45	-4.96
0.50	-91.04	238.61	-1.86	-49.18	0.04	-12.63	-9.12	-20.77	-22.09
0.75	-91.04	535.72	-4.20	-49.16	1.37	-27.96	-9.09	-49.00	-51.28

Table 3: Efficiency (%) of the $\tilde{X}^{CR}$, $\tilde{Y}^{R}$ and $\tilde{Y}^{CR}$ estimators (Poisson)

$\rho(\mathbf{y}.\mathbf{x})$	$(n_{s^\star}=500.n_s=5000)$			$(n_{s^\star}=5000.n_s=5000)$			$(n_{s^\star}=5000.n_s=500)$		
	$\mathbf{\tilde{X}^{CR}}\lvert\mathbf{\tilde{X}}$	$\mathbf{\tilde{Y}^{R}}\lvert\mathbf{\hat{Y}}$	$\mathbf{\tilde{Y}^{CR}}\lvert\mathbf{\hat{Y}}$	$\mathbf{\tilde{X}^{CR}}\lvert\mathbf{\tilde{X}}$	$\mathbf{\tilde{Y}^{R}}\lvert\mathbf{\hat{Y}}$	$\mathbf{\tilde{Y}^{CR}}\lvert\mathbf{\hat{Y}}$	$\mathbf{\tilde{X}^{CR}}\lvert\mathbf{\tilde{X}}$	$\mathbf{\tilde{Y}^{R}}\lvert\mathbf{\hat{Y}}$	$\mathbf{\tilde{Y}^{CR}}\lvert\mathbf{\hat{Y}}$
				Distribution 1: CV(y)=0.88. CV(x)=0.63					
0.25	-91.39	321.44	-3.97	-50.23	-14.40	-24.53	-9.61	-45.83	-44.82
0.50	-91.38	523.13	-5.55	-50.63	-8.36	-32.22	-9.01	-58.85	-58.52
0.75	-91.39	814.83	-6.61	-50.45	4.13	-40.32	-9.28	-72.00	-73.07
				Distribution 2: CV(y)=0.40. CV(x)=0.63					
0.25	-91.04	353.17	-4.86	-50.36	-23.80	-32.08	-9.00	-60.97	-59.03
0.50	-91.02	452.26	-5.45	-50.30	-23.17	-37.04	-9.20	-69.55	-67.78
0.75	-91.03	569.40	-6.23	-50.54	-20.70	-42.15	-9.19	-78.45	-77.01
				Distribution 3: CV(y)=0.88. CV(x)=1.41					
0.25	-91.24	40.04	-2.14	-49.50	-15.57	-10.81	-8.41	-21.32	-19.93
0.50	-91.24	108.60	-3.63	-49.46	-24.12	-19.14	-8.62	-37.42	-35.31
0.75	-91.23	242.72	-5.30	-49.85	-31.20	-30.08	-8.63	-58.70	-56.03
				Distribution 4: CV(y)=0.40. CV(x)=1.41					
0.25	-91.04	31.51	-2.05	-49.92	-16.92	-11.109	-8.63	-21.95	-20.44
0.50	-91.10	55.81	-2.74	-49.93	-22.17	-15.282	-8.61	-30.28	-28.31
0.75	-91.05	90.55	-3.66	-49.92	-28.06	-20.450	-8.55	-40.04	-37.58

Table 4: Efficiency (%) of the $\tilde{X}^{CO}$, $\tilde{Y}^{O}$ and $\tilde{Y}^{CO}$ estimators (Poisson)

$\rho(\mathbf{y},\mathbf{x})$	$(n_{s^\star}=500, n_s=5000)$			$(n_{s^\star}=5000, n_s=5000)$			$(n_{s^\star}=5000, n_s=500)$		
	$\tilde{\mathbf{X}}^{\mathbf{CO}}\|\tilde{\mathbf{X}}$	$\tilde{\mathbf{Y}}^{\mathbf{O}}\|\hat{\mathbf{Y}}$	$\tilde{\mathbf{Y}}^{\mathbf{CO}}\|\hat{\mathbf{Y}}$	$\tilde{\mathbf{X}}^{\mathbf{CO}}\|\tilde{\mathbf{X}}$	$\tilde{\mathbf{Y}}^{\mathbf{O}}\|\hat{\mathbf{Y}}$	$\tilde{\mathbf{Y}}^{\mathbf{CO}}\|\hat{\mathbf{Y}}$	$\tilde{\mathbf{X}}^{\mathbf{CO}}\|\tilde{\mathbf{X}}$	$\tilde{\mathbf{Y}}^{\mathbf{O}}\|\hat{\mathbf{Y}}$	$\tilde{\mathbf{Y}}^{\mathbf{CO}}\|\hat{\mathbf{Y}}$
				Distribution 1: CV(y)=0.88, CV(x)=0.63					
0.25	-91.40	508.91	-3.94	-50.22	1.90	-25.27	-9.62	-45.61	-46.22
0.50	-91.38	639.20	-5.55	-50.63	2.47	-32.36	-9.04	-58.09	-58.84
0.75	-91.39	789.78	-6.68	-50.45	1.75	-40.35	-9.32	-72.23	-73.07
				Distribution 2: CV(y)=0.40, CV(x)=0.63					
0.25	-91.04	654.50	-4.88	-50.36	2.65	-33.83	-9.03	-62.01	-62.65
0.50	-91.02	736.05	-5.40	-50.31	2.35	-38.37	-9.22	-69.79	-70.50
0.75	-91.03	820.91	-6.14	-50.54	2.62	-42.99	-9.22	-77.95	-78.78
				Distribution 3: CV(y)=0.88, CV(x)=1.41					
0.25	-91.25	305.56	-3.22	-49.51	0.95	-15.77	-8.41	-29.14	-29.39
0.50	-91.24	466.73	-4.62	-49.47	0.82	-24.25	-8.61	-44.49	-44.92
0.75	-91.24	668.64	-5.93	-49.85	1.93	-34.28	-8.60	-63.61	-64.23
				Distribution 4: CV(y)=0.40, CV(x)=1.41					
0.25	-91.05	344.18	-3.28	-49.91	1.80	-17.79	-8.70	-31.98	-32.48
0.50	-91.10	436.55	-3.97	-49.92	1.86	-22.54	-8.66	-40.79	-41.41
0.75	-91.06	538.09	-4.95	-49.92	2.00	-28.08	-8.62	-50.75	-51.42

5. Discussion

We have investigated the effect of calibration to estimated totals of auxiliary variables on the properties of the resulting pseudo-regression or pseudo-optimal estimator of the total of any survey variable. Obviously, since such an estimator of the total of an auxiliary variable is identical to the total estimated from the benchmark survey, there will be loss of precision if its variance has larger variance than the total estimated from the primary sample.

For any other survey variable, we have shown theoretically and empirically that the pseudo-regression estimator is grossly inefficient not only relative to the standard regression estimator but also relative to the basic Horvitz-Thompson estimator when the benchmark survey is of much smaller size than the primary survey; oddly this inefficiency increases as the correlation between this variable and the auxiliary variable increases. The pseudo-optimal estimator is also very inefficient, though less than the regression estimator. The inefficiency of the pseudo-regression estimator lessens as the size of the benchmark sample becomes larger than that of the primary survey. For such samples, and depending on the sampling design, the pseudo-regression estimator may be more efficient than the basic estimator; the same applies to the pseudo-optimal estimator. Notably, the pseudo-regression estimator is somewhat biased, but not the pseudo-optimal estimator.

We have also shown that the composite calibration procedure, feasible when micro-data from the benchmark survey are available, generates more efficient regression and optimal estimators. The composite regression estimator for the auxiliary variable is always more efficient than the basic estimator from the benchmark survey; the efficiency of this estimator reflects the relative sample size of the primary and benchmark survey. The efficiencies of the composite regression estimator and the composite optimal estimator are almost iden-

tical.

For any other survey variable, the composite regression estimator and the composite optimal estimator are always more efficient than the pseudo-regression and pseudo-optimal estimators, respectively. The efficiency increases as the correlation of the particular variable with the auxiliary variable used in the calibration increases, and as the sample size of the benchmark survey gets larger than that of the primary survey.

Besides producing consistent estimates for common variables between surveys, composite calibration produces more accurate estimates for any variable than the customary calibration procedure, and, more conveniently, valid estimates of their variances accounting for the variability of the random calibration totals.

Construction of composite optimal estimators involving two surveys with common variables with unknown totals, but not through a calibration procedure, and without the comparative analysis of Section 3, is presented also in Guandalini and Tillé (2015).

Finally, the calibration approach proposed in this article can be easily extended to more than two surveys with common variables.

REFERENCES

Andersson, P.G., and Thorburn, D. (2005). An optimal calibration distance leading to the optimal regression estimator. *Survey Methodology*, 31, 95–99.

Dever, J.A., and Valliant, R. (2010). A comparison of variance estimators for poststratification to estimated control totals. *Survey Methodology*, 36, 45–56.

Deville, J.C., and Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376–382.

Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28, 5–23.

Guandalini, A., and Tillé (2017). Design-based estimators calibrated on estimated totals from multiple surveys. *International Statistical Review*, 85, 250–269.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99, 1131–1139.

Merkouris, T. (2013). Composite calibration estimation integrating data from different surveys. Proceedings 59th ISI World Statistics Congress, Hong Kong, 205–210.

Montanari, G.E. (1987). Post-sampling efficient prediction in large scale surveys. *International Statistical Review*, 55, 191–202.

Montanari, G.E. (1998). On regression estimation of finite population means. *Survey Methodology*, 24, 69–77.

Opsomer, J.D., and Erciulescu, A.L. (2021). Replication variance estimation after sample-based calibration. *Survey Methodology*, 47, 265–277.

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153–165.

Särndal, C.E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33, 99–119.

Särndal, C.E., Swensson, B., and Wretman, J. H. (1992). *Model-Assisted Survey Sampling*, New York: Springer.