Journal of Official Statistics, Vol. 18, No. 2, 2002, pp 233–255

The Ten Cases of Auxiliary Information for Calibration in Two-Phase Sampling

Victor M. Estevao[1] *and Carl-Erik Särndal*[2]

Calibration is commonly used to produce estimation weights in sample surveys. Calibration weights satisfy a set of calibration equations that make use of the specified auxiliary information. In a two-phase design, the information used for calibration can take different forms. The case that we call *complete auxiliary information* arises when information is available at the level of the population for one set of auxiliary variables and at the lower level of the first-phase sample for another set of auxiliary variables. In practice, we may be restricted to a calibration on a subset of the complete auxiliary information, or we may decide to discard some of the complete information if no significant loss of efficiency occurs. We show that there are exactly nine different subsets of the complete information, for a total of ten different cases of auxiliary information. We propose one calibration estimator in each of these ten cases. In general, the more extensive the auxiliary information, the better the precision of the resulting estimates. However, there are sometimes surprising exceptions to this, as illustrated both by our theoretical results and by our simulation. We study the precision of the calibration estimators in the ten cases, both theoretically (by deriving the sum of the two variance components) and empirically (by repeated sampling from different types of populations). We suggest a simple approach to determine the best use of auxiliary information.

Key words: Design-based inference; linear regression representation; calibrated weights; regression residuals; variance estimation.

1. Introduction

A distinguishing feature of two-phase sampling designs is that auxiliary information may exist at two levels. Some information is at the level of the whole population and other information is at the level of the first-phase sample. We may use all, some or none of this information to obtain calibrated weights. These weights are then used to produce an estimate of the parameter of interest such as a population total. The variance of the calibration estimator depends on the level and amount of auxiliary information used in the calibration.

Two-phase sampling designs have attracted considerable attention in the recent literature. There are two reasons for this. First, the efficiency of two-phase designs has led to an increased use of them in statistical agencies such as Statistics Canada. In addition, two-phase sampling provides a simple mechanism for the handling of nonresponse. We select a sample and regard the respondents as the second-phase sample.

[1] Senior Methodologist, Statistics Canada, Ottawa, Ontario, K1A 0T6, Canada. E-mail: victor.estevao@statcan.ca
[2] Professor, Consultant, Ottawa, Canada. E-mail: carl.sarndal@rogers.com
Acknowledgment: We thank the Associate Editor and the two referees for their comments on earlier versions of this article. Their suggestions helped us improve the content and the presentation.

There is an interest in studying alternative uses of auxiliary information in two-phase designs: (i) it is important to identify the relevant auxiliary information for the calibration; (ii) the complete auxiliary information may not be available, so we often have to calibrate with a smaller set of information. We show that there are exactly ten different cases of calibration for a two-phase design, depending on whether we use all or part of the complete auxiliary information. These different quantities of information influence the precision (the variance) of the calibration estimator. We examine these issues from a theoretical perspective, by deriving the large sample variances of the ten cases, and from an empirical perspective through Monte Carlo simulation, in which repeated two-phase samples are drawn from six generated finite populations.

2. Calibration Estimation in Two-Phase Sampling

We consider a finite population of size N, denoted $U = \{1, \ldots, k, \ldots, N\}$. A first-phase sample s_1, of size n_1, is drawn from U with a design such that the sampling weight for unit k is $a_{1k} = 1/\pi_{1k}$, where $\pi_{1k} = P(k \in s_1)$ is the known first-phase inclusion probability of k. Some variables (although not the variable of interest) are observed for the first-phase sample units. A second-phase sample s, of size n, is drawn from s_1 with a second-phase design such that the (conditional) sampling weight for unit k is $a_{2k} = 1/\pi_{2k}$, where $\pi_{2k} = P(k \in s \mid s_1)$ is the (conditional) inclusion probability of k, given s_1. The total sampling weight of unit k, given by $a_k = a_{1k}a_{2k}$, is called the *design weight*.

The variable of interest is denoted y; its value for unit k is y_k. The target of estimation is the population total $Y = \sum_U y_k$. For simplicity, we write $\sum_{k \in A}$ as $\sum_A$ for any $A \subseteq U$. The y-variable is observed only for the second-phase sample units, so the available y-data are $\{y_k: k \in s\}$. The two-phase double expansion estimator, given by $\hat{Y}_{DE} = \sum_s a_{1k}a_{2k}y_k$, is unbiased for Y but makes no use of auxiliary information. It is however a natural point of reference with which to compare the usually more efficient alternatives that follow.

Calibration is used to modify the design weights a_k subject to constraints called *calibration equations*. In this article, we produce a set of *calibrated weights* $\{w_k: k \in s\}$ of the form $w_k = a_k g_k$, where g_k is the *weight adjustment factor* for unit k. As shown in Section 5, it is important that g_k be close to 1 for all units in order to obtain an approximately unbiased estimator and to permit the estimation of variance. We estimate Y by applying the weight w_k to the observed value y_k. Summing over the units in the second-phase sample, we obtain the *two-phase calibration estimator*

$$\hat{Y} = \sum_s w_k y_k \tag{2.1}$$

Estimators for two-phase designs can also be constructed using a *regression approach*. Särndal and Swensson (1987) and Särndal, Swensson, and Wretman (1992) examine different regression estimators for two-phase designs. Armstrong and St-Jean (1994) applied regression estimation in a Statistics Canada survey with a two-phase design. Binder (1996) gives a useful linearization technique to obtain the approximate variance of nonlinear estimators. Dupont (1995) studied regression and calibration and the relation between them. Axelson (2000) discusses alternative approaches to variance estimation. Lundström (1997) describes techniques for calibration to handle nonresponse.

The usual approach to computing calibrated weights is by *distance minimization*, which

requires the specification of a distance function. This is discussed, mostly in connection with one-phase designs, in Huang and Fuller (1978), Alexander (1987), Bankier (1989), Deville and Särndal (1992), Deville, Särndal, and Sautory (1993), and Singh and Mohl (1996). As some of these references show, differences are often negligible between the estimates produced by different distance measures. Another approach to calibration is the functional form method given by Estevao and Särndal (2000) in which the calibrated weights are given an explicit functional form. In this article, we examine the differences between calibration estimators in the ten different cases of auxiliary information, not between calibration estimators within each case. The reason is that we do not always have control over the amount of auxiliary information available for calibration, but we know how to compute calibrated weights within each case. We derive an efficient calibration estimator for each of the ten cases, using least-squares minimization in one or two steps to create the calibrated weights w_k.

Consider two auxiliary vectors denoted $\mathbf{x}_1$ and $\mathbf{x}_2$ with $J_1 \geqslant 1$ and $J_2 \geqslant 1$ auxiliary variables, respectively. The values of $\mathbf{x}_1$ and $\mathbf{x}_2$ for unit k are denoted by $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$. We assume that we have the following auxiliary information.

- The vector total $\sum_U \mathbf{x}_{1k}$ is known.
- $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$ are known vector values for every $k \in s_1$.

We refer to this as the *complete auxiliary information*. The calibration process may use all or part of it to produce the calibrated weights $\{w_k\colon k \in s\}$ and the calibration estimator $\hat{Y} = \sum_s w_k y_k$. Since each calibration equation always involves two different levels of information, it is useful to present the auxiliary information by level as follows:

- *At the level of the population U*: The vector total $\sum_U \mathbf{x}_{1k}$ is known.
- *At the level of the first-phase sample* s_1: $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$ are known for every $k \in s_1$.
- *At the level of the second-phase sample s*: $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$ are known for every $k \in s$.

This article is arranged as follows. In Section 3, we show that there are exactly ten ways of specifying the calibration equations. Each of these cases is a different way of using the complete auxiliary information for calibration. Within each case, we present the calibration equations and define one calibration estimator through a simple method of calculating an efficient set of weights w_k. The calibration equations and the calculation of the calibration weights are discussed in Section 4. The bias and approximate variance of the estimator are derived in Section 5 and the estimation of variance is shown in Section 6. In Section 7 we describe an empirical simulation and examine its results. In Section 8 we provide a simple recommendation to produce the most efficient estimator among the possible cases.

3. The Ten Cases of Auxiliary Information

Consider the complete auxiliary information given in Section 2. It is possible to make use of it in different ways. At the two extremes, we use either all or none of this information. When no information is used, we obtain the double expansion estimator. There are also eight intermediate cases, which use some auxiliary information. These cases merit attention because of the following:

(i) In some situations, there is little loss in efficiency when we ignore some of the auxiliary information in the calibration. It is even possible to obtain a calibrated estimator whose variance is less than that based on the complete information. This is illustrated in the simulation results of Section 7.

(ii) We do not always have complete auxiliary information so we have to make do with what is available. It is important to see how this limitation affects the variance of the estimator and how the latter variance compares to the variance of the estimators in the two extreme situations.

We enumerate the ten possible cases by coding them using a sextuplet *zzz/zzz*, where each position z is either 1 or 0 to identify whether or not we use the corresponding auxiliary information. The first three positions indicate the use of information on $\mathbf{x}_1$ in the calibration. Similarly, the last three positions tell us how we use information on $\mathbf{x}_2$. The first and fourth positions indicate the use ($z=1$) or non-use ($z=0$) of information at level U, the second and fifth at level s_1, and the third and sixth at level s. Each calibration equation involves two sums at different levels. We always calibrate from a *lower level* to a *higher level*. Therefore, the sum represented by the higher level must be known. The sum at the lower level involves the weights to be determined. As a rule, the lower level is shown on the left of the equation and the higher level is shown on the right. Furthermore, if we require the first-phase calibrated weights w_{1k} in the calibration equations then we must calculate these first, and use them to obtain the weights w_k.

We illustrate this coding structure for the case where we use the complete auxiliary information. This is Case A1 in Table 1. First, we calibrate on $\mathbf{x}_1$ from s_1 to U, to produce the weights w_{1k}. At this point, the code is $11z/0zz$. Then, we calibrate on $\mathbf{x}_1$ and $\mathbf{x}_2$ from s to

Table 1. The ten different cases of auxiliary information for calibration

Case	Code	Calibration features and sequence	Calibration equations
A	$11z/0zz$	Calibration on $\mathbf{x}_1$ from s_1 to U	
A1	111/011	$\mathbf{x}_1$ from s_1 to U to obtain w_{1k}	$\sum_{s_1} w_{1k}\mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$
		then $\mathbf{x} = (\mathbf{x}_1', \mathbf{x}_2')'$ from s to s_1 to obtain w_k	$\sum_s w_k\mathbf{x}_k = \sum_{s_1} w_{1k}\mathbf{x}_k$
A2	111/000	$\mathbf{x}_1$ from s_1 to U to obtain w_{1k}	$\sum_{s_1} w_{1k}\mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$
		then $\mathbf{x}_1$ from s to s_1 to obtain w_k	$\sum_s w_k\mathbf{x}_{1k} = \sum_{s_1} w_{1k}\mathbf{x}_{1k}$
A3	110/011	$\mathbf{x}_1$ from s_1 to U to obtain w_{1k}	$\sum_{s_1} w_{1k}\mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$
		then $\mathbf{x}_2$ from s to s_1 to obtain w_k	$\sum_s w_k\mathbf{x}_{2k} = \sum_{s_1} w_{1k}\mathbf{x}_{2k}$
A4	110/000	$\mathbf{x}_1$ from s_1 to U to obtain w_{1k}, then $w_k = w_{1k}a_{2k}$	$\sum_{s_1} w_{1k}\mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$
B	$101/0zz$	Calibration on $\mathbf{x}_1$ from s to U	
B1	101/011	$\mathbf{x}_1$ from s to U and $\mathbf{x}_2$ from s to s_1 to obtain w_k	$\sum_s w_k\mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$
			$\sum_s w_k\mathbf{x}_{2k} = \sum_{s_1} a_{1k}\mathbf{x}_{2k}$
B2	101/000	$\mathbf{x}_1$ from s to U to obtain w_k	$\sum_s w_k\mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$
C	$0zz/0zz$	No calibration on $\mathbf{x}_1$ to level U	
C1	011/011	$\mathbf{x} = (\mathbf{x}_1', \mathbf{x}_2')'$ from s to s_1 to obtain w_k	$\sum_s w_k\mathbf{x}_k = \sum_{s_1} a_{1k}\mathbf{x}_k$
C2	011/000	$\mathbf{x}_1$ from s to s_1 to obtain w_k	$\sum_s w_k\mathbf{x}_{1k} = \sum_{s_1} a_{1k}\mathbf{x}_{1k}$
C3	000/011	$\mathbf{x}_2$ from s to s_1 to obtain w_k	$\sum_s w_k\mathbf{x}_{2k} = \sum_{s_1} a_{1k}\mathbf{x}_{2k}$
C4	000/000	(none)	(none)

s_1, to produce the weights w_k. This results in the code 111/011 for Case A1. It follows from our description of complete auxiliary information that a ''1'' is possible in all six positions of the code except in position four, which is always ''0'' because we cannot calibrate on $\mathbf{x}_2$ to level U.

Any calibration using less than the complete information is also uniquely coded by this notation. There are five possibilities for the first three positions of the code which specify the calibration on $\mathbf{x}_1$: 111/ (from s_1 to U and from s to s_1); 110/ (from s_1 to U only); 101/ (from s directly to U, bypassing s_1); 011/ (from s to s_1 only); 000/ (no calibration on $\mathbf{x}_1$). For the last three positions of the code, specifying the calibration on $\mathbf{x}_2$, there are two possibilities: /011 (from s to s_1); /000 (no calibration on $\mathbf{x}_2$). This results in $5 \times 2 = 10$ possible calibration cases. They are listed in Table 1. We divide the ten cases into three general categories: Cases A, B, and C. This classification provides a more convenient grouping for examining the properties of the estimators. The list in Table 1 progresses from ''complete information'' (Case A1, at the top) to ''no information'' (Case C4, at the bottom). It is not possible to give a completely ordered progression, where a case higher in the list has more information than one lower in the list. For example, we conclude that A1 has more information than A2, A3, B1, or C1, but among these, we cannot say that one has more auxiliary information than the others.

We examine all ten cases. Some cases, such as A1 and B1, show subtle differences in the use of auxiliary information. Unlike A1, B1 does not require the $\mathbf{x}_{1k}$ values over the first-phase sample. Case B1 only requires that we observe $\mathbf{x}_{1k}$ for units in the second-phase sample, and that we know the population vector total $\sum_U \mathbf{x}_{1k}$ from a census or other reliable data source outside the survey. This can happen in practice. The lack of the individual values of $\mathbf{x}_{1k}$ for $k \in s_1$ forces us to change the calibration equations and this leads to an estimator different from that of A1. Although Case B1 requires less information than A1, this does not always lead to a less efficient estimator. There are situations where B1 produces a calibration estimator with smaller variance than that of A1. This and other results are shown in the simulation of Section 7.

4. The Calibration Equations for the Ten Cases

The calibration equations for all of the ten cases are determined by the sextuplet code. These equations are shown in Table 1. Case A in this table is a generic representation covering the four Cases A1, A2, A3, and A4. The code type is 11z/0zz, indicating a calibration on $\mathbf{x}_1$ from s_1 to U, to obtain the intermediate weights, $\{w_{1k}: k \in s_1\}$. In Cases A1, A2, and A3, this is followed by a calibration from s to s_1 to produce the final weights $\{w_k: k \in s\}$. The calibration equations for Case A are

$$\begin{aligned} \sum_{s_1} w_{1k}\mathbf{x}_{1k} &= \sum_U \mathbf{x}_{1k} \\ \sum_s w_k\mathbf{x}_k &= \sum_{s_1} w_{1k}\mathbf{x}_k \end{aligned} \tag{4.1}$$

where $\mathbf{x}_k$ is one of the following, corresponding to A1, A2, A3, and A4, respectively:

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_{1k} \\ \mathbf{x}_{2k} \end{pmatrix}; \quad \mathbf{x}_k = \mathbf{x}_{1k}; \quad \mathbf{x}_k = \mathbf{x}_{2k}; \quad \mathbf{x}_k = \phi \text{ for all } k \tag{4.2}$$

The calibrated weights are computed by minimizing an objective function in each of the two steps, as described below. In the first step, the first-phase design weights a_{1k} are adjusted to obtain calibrated weights $w_{1k} = a_{1k}g_{1k}$ satisfying the first calibration equation in (4.1). Here, g_{1k} is a first-phase weight adjustment factor for each unit $k \in s_1$; it does not depend on the second-phase sample. The weights w_{1k} are then used to obtain $\sum_{s_1} w_{1k}\mathbf{x}_k$ for the second calibration equation. Then, the second-phase design weights $a_{1k}a_{2k}$ are adjusted to obtain calibrated weights $w_k = a_{1k}a_{2k}g_k$ satisfying the second equation in (4.1). Here, g_k is the weight adjustment factor for each unit $k \in s$. In A4, there is no second calibration equation since $\mathbf{x}_k = \phi$. The weights w_k are obtained from w_{1k} by defining $w_k = w_{1k}a_{2k}$.

Cases B and C do not involve calibration from s_1 to U. Consequently, there is no calculation of first-phase weights w_{1k}. Case B covers B1 and B2. It has code 101/0*zz*, denoting a calibration on $\mathbf{x}_1$ directly from s to U. Thus, the first-phase sampling weights a_{1k} remain unchanged in sums over s_1. The calibration equations are therefore

$$\begin{aligned} \sum_s w_k\mathbf{x}_{1k} &= \sum_U \mathbf{x}_{1k} \\ \sum_s w_k\mathbf{x}_k &= \sum_{s_1} a_{1k}\mathbf{x}_k \end{aligned} \tag{4.3}$$

with $\mathbf{x}_k = \mathbf{x}_{2k}$ (B1) or $\mathbf{x}_k = \phi$ (B2). The second equation disappears for B2.

Case C in Table 1 covers four cases. It has the code type 0*zz*/0*zz*, indicating no calibration to U. Therefore, the only calibration equation is

$$\sum_s w_k\mathbf{x}_k = \sum_{s_1} a_{1k}\mathbf{x}_k \tag{4.4}$$

where $\mathbf{x}_k$ is one of the four alternatives given in (4.2), leading to C1, C2, C3, and C4, respectively. Case C4, with $\mathbf{x}_k = \phi$, involves no calibration; the final weights w_k are simply $w_k = a_k$, the sampling design weights for all $k \in s$.

It follows that Cases B and C (except C4) only require a single adjustment of the design weights $a_{1k}a_{2k}$ to arrive at the final weights $w_k = a_{1k}a_{2k}g_k$ for $k \in s$. The value of g_k for these cases is generally different from that obtained for Case A. The weight computation for Cases B and C involves the minimization of one objective function.

The calibration weights w_{1k} and w_k for A1, A2, and A3 were obtained by the following approach. We express each weight as the corresponding design weight multiplied by a weight adjustment factor. Thus, we write

$$\begin{aligned} w_{1k} &= a_{1k}g_{1k} \quad \text{for} \quad k \in s_1 \\ w_k &= a_{1k}a_{2k}g_k \quad \text{for} \quad k \in s \end{aligned} \tag{4.5}$$

Substituting these terms into the calibration equations for these three cases, we obtain a set of equations in terms of the unknowns g_{1k} and g_k. For example, the Calibration equations (4.1) for A1 become

$$\begin{aligned} \sum_{s_1} a_{1k}g_{1k}\mathbf{x}_{1k} &= \sum_U \mathbf{x}_{1k} \\ \sum_s a_{1k}a_{2k}g_k\mathbf{x}_k &= \sum_{s_1} a_{1k}g_{1k}\mathbf{x}_k \end{aligned} \tag{4.6}$$

There are many solutions for g_{1k} and g_k. To have a bias close to zero and to permit the estimation of variance, we need to have $g_{1k} \doteq 1$ for all $k \in s_1$, and $g_k \doteq 1$ for all $k \in s$ or equivalently, $w_{1k} \doteq a_{1k}$ for all $k \in s_1$, and $w_k \doteq a_{1k}a_{2k}$ for all $k \in s$. One way to do this is by the following two-step procedure. Step 1: Determine the weights $w_{1k} = a_{1k}g_{1k}$ as the solution to the weighted least squares minimization problem given by

$$\begin{aligned} &\text{Min} \sum_{s_1} \frac{(w_{1k} - a_{1k})^2}{a_{1k}} \\ &\text{subject to} \sum_{s_1} w_{1k}\mathbf{x}_{1k} = \sum_{U} \mathbf{x}_{1k} \end{aligned} \tag{4.7}$$

The solution is given by the weights

$$w_{1k} = a_{1k} + \left(\sum_{U} \mathbf{x}_{1k} - \sum_{s_1} a_{1k}\mathbf{x}_{1k}\right)' \left(\sum_{s_1} a_{1k}\mathbf{x}_{1k}\mathbf{x}'_{1k}\right)^{-1} a_{1k}\mathbf{x}_{1k} \tag{4.8}$$

We then use these weights to produce $\sum_{s_1} w_{1k}\mathbf{x}_k$ for the right-hand side of the second equation in (4.6). Step 2: The weights $w_k = a_{1k}a_{2k}g_k$ are obtained as the solution to the minimization problem given by

$$\begin{aligned} &\text{Min} \sum_{s} \frac{(w_k - a_{1k}a_{2k})^2}{a_{1k}a_{2k}} \\ &\text{subject to} \sum_{s} w_k\mathbf{x}_k = \sum_{s_1} w_{1k}\mathbf{x}_k \end{aligned} \tag{4.9}$$

This gives the weights

$$w_k = a_{1k}a_{2k} + \left(\sum_{s_1} w_{1k}\mathbf{x}_k - \sum_{s} a_{1k}a_{2k}\mathbf{x}_k\right)' \left(\sum_{s} a_{1k}a_{2k}\mathbf{x}_k\mathbf{x}'_k\right)^{-1} a_{1k}a_{2k}\mathbf{x}_k \tag{4.10}$$

Case A4 only requires a single step calibration since there is no second calibration equation. Once the weights w_{1k} are obtained, we simply define $w_k = w_{1k}a_{2k}$. The estimators in Cases B and C do not involve the weights w_{1k}, only the final weights $w_k = a_{1k}a_{2k}g_k$. These are obtained in one step by minimizing the objective function $\sum_s (w_k - a_{1k}a_{2k})^2 / a_{1k}a_{2k}$ subject to the corresponding calibration equations. It is interesting to note that the use of the least squares function in the minimization problems produces the same weights w_k, and consequently the same estimator for A2 and B2. In general, other objective functions may produce different weights for Cases A2 and B2.

5. Bias and Variance of the Two-Phase Calibration Estimators

In this section, we derive the bias and the approximate variance of the calibration estimator for the ten cases of auxiliary information. Table 2 provides a summary of the important properties of the estimators. The approximate variances are given in terms of regression residuals and generally indicate (although not always unequivocally) which estimators are expected to have the smallest variance. Calibration is involved in all cases except C4, which gives the unbiased double expansion estimator. This estimator has a well-known exact expression for the variance. In every other case, the resulting calibration

estimator is approximately unbiased and the variance expression is approximate. The residuals in the two terms of the approximate variance are shown in Table 2.

The derivation of the approximate variances requires the use of one or two *least squares linear regression representations* over all units $k \in U$. The calibration equations determine the representations appropriate for any given case. These representations do not require the assumptions associated with traditional methods of regression model fitting.

For the general form of Case A, the representations are obtained by a sequential procedure which identifies the auxiliary variables used to obtain w_k and w_{1k}. For the first linear representation, we look for the auxiliary variables in the calibration equation used to determine w_k and we express y_k as a linear regression of these variables. From the calibration equations for Case A in Table 2, we find that these variables are given by $\mathbf{x}_k$. Therefore, we write $y_k = \mathbf{x}'_k\mathbf{B}_{(y;\mathbf{x})} + e_{k(y;\mathbf{x})}$ for $k \in U$. The term $\mathbf{B}_{(y;\mathbf{x})}$ is a population regression coefficient and $e_{k(y;\mathbf{x})}$ is the population residual for $k \in U$. We use the subscript $(y;\mathbf{x})$ to indicate the regression of y_k on $\mathbf{x}_k$. The coefficient $\mathbf{B}_{(y;\mathbf{x})}$ is defined by the ordinary least squares (OLS) minimization of $\sum_U (y_k - \mathbf{x}'_k\mathbf{B}_{(y;\mathbf{x})})^2$. Therefore, the residuals satisfy the normal equation

$$\sum_U \mathbf{x}_k(y_k - \mathbf{x}'_k\mathbf{B}_{(y;\mathbf{x})}) = \sum_U \mathbf{x}_k e_{k(y;\mathbf{x})} = \mathbf{0} \tag{5.1}$$

For the second linear representation, we take the predicted values $\mathbf{x}'_k\mathbf{B}_{(y;\mathbf{x})}$ and express them as a linear regression of the auxiliary variables in the calibration equation used to determine w_{1k}. These variables are given by $\mathbf{x}_{1k}$. Therefore, we write $\mathbf{x}'_k\mathbf{B}_{(y;\mathbf{x})} = \mathbf{x}'_{1k}\mathbf{B}_{(\mathbf{xB};\mathbf{x}_1)} + e_{k(\mathbf{xB};\mathbf{x}_1)}$ for $k \in U$. The term $\mathbf{B}_{(\mathbf{xB};\mathbf{x}_1)}$ is the population regression coefficient (of dimension J_1) for this representation and $e_{k(\mathbf{xB};\mathbf{x}_1)}$ is the corresponding population residual for $k \in U$. The subscript $(\mathbf{xB};\mathbf{x}_1)$ indicates the regression of $\mathbf{x}'_k\mathbf{B}_{(y;\mathbf{x})}$ on $\mathbf{x}_{1k}$. The coefficient $\mathbf{B}_{(\mathbf{xB};\mathbf{x}_1)}$ is the solution to the OLS minimization of $\sum_U (\mathbf{x}'_k\mathbf{B}_{(y;\mathbf{x})} - \mathbf{x}'_{1k}\mathbf{B}_{(\mathbf{xB};\mathbf{x}_1)})^2$. Therefore, the residuals $e_{k(\mathbf{xB};\mathbf{x}_1)}$ satisfy the normal equation

$$\sum_U \mathbf{x}_{1k}(\mathbf{x}'_k\mathbf{B}_{(y;\mathbf{x})} - \mathbf{x}'_{1k}\mathbf{B}_{(\mathbf{xB};\mathbf{x}_1)}) = \sum_U \mathbf{x}_{1k} e_{k(\mathbf{xB};\mathbf{x}_1)} = \mathbf{0} \tag{5.2}$$

In summary, the representations for Case A are

$$\begin{aligned} y_k &= \mathbf{x}'_k\mathbf{B}_{(y;\mathbf{x})} + e_{k(y;\mathbf{x})} \\ \mathbf{x}'_k\mathbf{B}_{(y;\mathbf{x})} &= \mathbf{x}'_{1k}\mathbf{B}_{(\mathbf{xB};\mathbf{x}_1)} + e_{k(\mathbf{xB};\mathbf{x}_1)} \end{aligned} \tag{5.3}$$

where the first three forms of $\mathbf{x}_k$ in (4.2) agree with the calibration equations for A1, A2, and A3, respectively. For A4, there is no calibration from s to s_1 since $\mathbf{x}_k = \phi$. Consequently, (5.3) does not provide a proper representation. The appropriate representation for A4 is the single equation $y_k = \mathbf{x}'_{1k}\mathbf{B}_{(y;\mathbf{x}_1)} + e_{k(y;\mathbf{x}_1)}$. A similar approach can be used to obtain the representations for Cases B and C. In Case B, we express y_k as a linear regression of the auxiliary variables in the calibration equations used to determine w_k. Thus, we obtain $y_k = (\mathbf{x}'_{1k}, \mathbf{x}'_k)\mathbf{B}_{(y;\mathbf{x}_1,\mathbf{x})} + e_{k(y;\mathbf{x}_1,\mathbf{x})}$ with $\mathbf{x}_k = \mathbf{x}_{2k}$ (for B1) or $\mathbf{x}_k = \phi$ (for B2). The components of $\mathbf{B}_{(y;\mathbf{x}_1,\mathbf{x})}$ associated with $\mathbf{x}_1$ and $\mathbf{x}$ are given by $\mathbf{B}^{(1)}_{(y;\mathbf{x}_1,\mathbf{x})}$ and $\mathbf{B}^{(2)}_{(y;\mathbf{x}_1,\mathbf{x})}$. The representations for Case C are obtained in a similar manner. All representations are shown in Table 2. Using (5.1), (5.2), and (5.3), we obtain the following properties for Case A:

Table 2. Properties of the calibration estimators for the ten cases of auxiliary information

Case	Auxiliary $\mathbf{x}$	Calibration equations	Linear representation	Residuals in approx. variance (5.13)	
				e_{1k}	e_{2k}
A	$\mathbf{x}_k$	$\sum_{s_1} w_{1k}\mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$ $\sum_s w_k\mathbf{x}_k = \sum_{s_1} w_{1k}\mathbf{x}_k$	$y_k = \mathbf{x}_k'\mathbf{B}_{(y;\mathbf{x})} + e_{k(y;\mathbf{x})}$ $\mathbf{x}_k'\mathbf{B}_{(y;\mathbf{x})} = \mathbf{x}_{1k}'\mathbf{B}_{(\mathbf{xB};\mathbf{x}_1)} + e_{k(\mathbf{xB};\mathbf{x}_1)}$	$y_k - \mathbf{x}_{1k}'\mathbf{B}_{(\mathbf{xB};\mathbf{x}_1)}$	$y_k - \mathbf{x}_k'\mathbf{B}_{(y;\mathbf{x})}$
A1	$\mathbf{x}_k = \begin{pmatrix}\mathbf{x}_{1k}\\ \mathbf{x}_{2k}\end{pmatrix}$	$\sum_{s_1} w_{1k}\mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$ $\sum_s w_k\mathbf{x}_k = \sum_{s_1} w_{1k}\mathbf{x}_k$	$y_k = \mathbf{x}_k'\mathbf{B}_{(y;\mathbf{x})} + e_{k(y;\mathbf{x})}$ $\mathbf{x}_k'\mathbf{B}_{(y;\mathbf{x})} = \mathbf{x}'_{1k}\mathbf{B}_{(\mathbf{xB};\mathbf{x}_1)} + e_{k(\mathbf{xB};\mathbf{x}_1)}$	$y_k - \mathbf{x}'_{1k}\mathbf{B}_{(y;\mathbf{x}_1)}$	$y_k - (\mathbf{x}_{1k}', \mathbf{x}_{2k}')\mathbf{B}_{(y;\mathbf{x}_1,\mathbf{x}_2)}$
A2	$\mathbf{x}_k = \mathbf{x}_{1k}$	$\sum_{s_1} w_{1k}\mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$ $\sum_s w_k\mathbf{x}_{1k} = \sum_{s_1} w_{1k}\mathbf{x}_{1k}$	$y_k = \mathbf{x}_{1k}'\mathbf{B}_{(y;\mathbf{x}_1)} + e_{k(y;\mathbf{x}_1)}$	$y_k - \mathbf{x}_{1k}'\mathbf{B}_{(y;\mathbf{x}_1)}$	$y_k - \mathbf{x}_{1k}'\mathbf{B}_{(y;\mathbf{x}_1)}$
A3	$\mathbf{x}_k = \mathbf{x}_{2k}$	$\sum_{s_1} w_{1k}\mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$ $\sum_s w_k\mathbf{x}_{2k} = \sum_{s_1} w_{1k}\mathbf{x}_{2k}$	$y_k = \mathbf{x}_{2k}'\mathbf{B}_{(y;\mathbf{x}_2)} + e_{k(y;\mathbf{x}_2)}$ $\mathbf{x}_{2k}'\mathbf{B}_{(y;\mathbf{x}_2)} = \mathbf{x}_{1k}'\mathbf{B}_{(\mathbf{x}_2\mathbf{B};\mathbf{x}_1)} + e_{k(\mathbf{x}_2\mathbf{B};\mathbf{x}_1)}$	$y_k - \mathbf{x}_{1k}'\mathbf{B}_{(\mathbf{x}_2\mathbf{B};\mathbf{x}_1)}$	$y_k - \mathbf{x}_{2k}'\mathbf{B}_{(y;\mathbf{x}_2)}$
A4	$\mathbf{x}_k = \phi$	$\sum_{s_1} w_{1k}\mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$ (then $w_k = w_{1k}a_{2k}$)	$y_k = \mathbf{x}_{1k}'\mathbf{B}_{(y;\mathbf{x}_1)} + e_{k(y;\mathbf{x}_1)}$	$y_k - \mathbf{x}_{1k}'\mathbf{B}_{(y;\mathbf{x}_1)}$	y_k
B	$\mathbf{x}_k$	$\sum_s w_k\mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$ $\sum_s w_k\mathbf{x}_k = \sum_{s_1} a_{1k}\mathbf{x}_k$	$y_k = (\mathbf{x}_{1k}', \mathbf{x}_k')\mathbf{B}_{(y;\mathbf{x}_1,\mathbf{x})} + e_{k(y;\mathbf{x}_1,\mathbf{x})}$	$y_k - \mathbf{x}_{1k}'\mathbf{B}^{(1)}_{(y;\mathbf{x}_1,\mathbf{x})}$	$y_k - (\mathbf{x}_{1k}', \mathbf{x}_k')\mathbf{B}_{(y;\mathbf{x}_1,\mathbf{x})}$
B1	$\mathbf{x}_k = \mathbf{x}_{2k}$	$\sum_s w_k\mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$ $\sum_s w_k\mathbf{x}_{2k} = \sum_{s_1} a_{1k}\mathbf{x}_{2k}$	$y_k = (\mathbf{x}_{1k}', \mathbf{x}_{2k}')\mathbf{B}_{(y;\mathbf{x}_1,\mathbf{x}_2)} + e_{k(y;\mathbf{x}_1,\mathbf{x}_2)}$	$y_k - \mathbf{x}_{1k}'\mathbf{B}^{(1)}_{(y;\mathbf{x}_1,\mathbf{x}_2)}$	$y_k - (\mathbf{x}_{1k}', \mathbf{x}_{2k}')\mathbf{B}_{(y;\mathbf{x}_1,\mathbf{x}_2)}$
B2	$\mathbf{x}_k = \phi$	$\sum_s w_k\mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$	$y_k = \mathbf{x}_{1k}'\mathbf{B}_{(y;\mathbf{x}_1)} + e_{k(y;\mathbf{x}_1)}$	$y_k - \mathbf{x}_{1k}'\mathbf{B}_{(y;\mathbf{x}_1)}$	$y_k - \mathbf{x}_{1k}'\mathbf{B}_{(y;\mathbf{x}_1)}$
C	$\mathbf{x}_k$	$\sum_s w_k\mathbf{x}_k = \sum_{s_1} a_{1k}\mathbf{x}_k$	$y_k = \mathbf{x}_k'\mathbf{B}_{(y;\mathbf{x})} + e_{k(y;\mathbf{x})}$	y_k	$y_k - \mathbf{x}_k'\mathbf{B}_{(y;\mathbf{x})}$
C1	$\mathbf{x}_k = \begin{pmatrix}\mathbf{x}_{1k}\\ \mathbf{x}_{2k}\end{pmatrix}$	$\sum_s w_k\mathbf{x}_{1k} = \sum_{s_1} a_{1k}\mathbf{x}_{1k}$ $\sum_s w_k\mathbf{x}_{2k} = \sum_{s_1} a_{1k}\mathbf{x}_{2k}$	$y_k = (\mathbf{x}_{1k}', \mathbf{x}_{2k}')\mathbf{B}_{(y;\mathbf{x}_1,\mathbf{x}_2)} + e_{k(y;\mathbf{x}_1,\mathbf{x}_2)}$	y_k	$y_k - (\mathbf{x}_{1k}', \mathbf{x}_{2k}')\mathbf{B}_{(y;\mathbf{x}_1,\mathbf{x}_2)}$
C2	$\mathbf{x}_k = \mathbf{x}_{1k}$	$\sum_s w_k\mathbf{x}_{1k} = \sum_{s_1} a_{1k}\mathbf{x}_{1k}$	$y_k = \mathbf{x}_{1k}'\mathbf{B}_{(y;\mathbf{x}_1)} + e_{k(y;\mathbf{x}_1)}$	y_k	$y_k - \mathbf{x}_{1k}'\mathbf{B}_{(y;\mathbf{x}_1)}$
C3	$\mathbf{x}_k = \mathbf{x}_{2k}$	$\sum_s w_k\mathbf{x}_{2k} = \sum_{s_1} a_{1k}\mathbf{x}_{2k}$	$y_k = \mathbf{x}_{2k}'\mathbf{B}_{(y;\mathbf{x}_2)} + e_{k(y;\mathbf{x}_2)}$	y_k	$y_k - \mathbf{x}_{2k}'\mathbf{B}_{(y;\mathbf{x}_2)}$
C4	$\mathbf{x}_k = \phi$	(none)	(none)	y_k	y_k

Property 1 If $\mathbf{x}_1$ is contained in $\mathbf{x}$ ($\mathbf{x}_1 \subseteq \mathbf{x}$) then:

(i) $\sum_U e_{k(y;\mathbf{x})} e_{k(\mathbf{xB};\mathbf{x}_1)} = 0$

(ii) $e_{k(y;\mathbf{x})} + e_{k(\mathbf{xB};\mathbf{x}_1)} = e_{k(y;\mathbf{x}_1)}$ for every $k \in U$, where the $e_{k(y;\mathbf{x}_1)}$ are the residuals of the OLS representation $y_k = \mathbf{x}'_{1k}\mathbf{B}_{(y;\mathbf{x}_1)} + e_{k(y;\mathbf{x}_1)}$

(iii) $\mathbf{B}_{(\mathbf{xB};\mathbf{x}_1)} = \mathbf{B}_{(y;\mathbf{x}_1)}$

Property 2 If $\mathbf{x}$ is contained in $\mathbf{x}_1$ ($\mathbf{x} \subseteq \mathbf{x}_1$) then $e_{k(\mathbf{xB};\mathbf{x}_1)} = 0$ for all $k \in U$.

Property 1 holds for A1. Both properties are true for A2. We now use the linear representations to derive the bias and approximate variance of the calibration estimators in Case A. Using (5.3) and the calibration equations (4.1), we find that the calibration estimator (2.1) for Case A can be written as

$$\hat{Y}_{\mathrm{A}} = \sum_U \mathbf{x}'_{1k}\mathbf{B}_{(\mathbf{xB};\mathbf{x}_1)} + \sum_{s_1} w_{1k} e_{k(\mathbf{xB};\mathbf{x}_1)} + \sum_s w_k e_{k(y;\mathbf{x})} \tag{5.4}$$

where w_{1k} and w_k are given by (4.8) and (4.10). Substituting for w_{1k} and w_k and simplifying we obtain

$$\begin{aligned}\hat{Y}_{\mathrm{A}} = &\sum_U \mathbf{x}'_{1k}\mathbf{B}_{(\mathbf{xB};\mathbf{x}_1)} + \sum_{s_1} a_{1k} e_{k(\mathbf{xB};\mathbf{x}_1)} + \sum_s a_{1k}a_{2k} e_{k(y;\mathbf{x})} \\ &+ \left(\sum_U \mathbf{x}_{1k} - \sum_{s_1} a_{1k}\mathbf{x}_{1k}\right)' (\hat{\mathbf{B}}_{(\mathbf{xB};\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{xB};\mathbf{x}_1)}) \\ &+ \left(\sum_{s_1} a_{1k}\mathbf{x}_k - \sum_s a_{1k}a_{2k}\mathbf{x}_k\right)' (\hat{\mathbf{B}}_{(y;\mathbf{x})} - \mathbf{B}_{(y;\mathbf{x})})\end{aligned} \tag{5.5}$$

where

$$\begin{aligned}\hat{\mathbf{B}}_{(y;\mathbf{x})} &= \left(\sum_s a_{1k}a_{2k}\mathbf{x}_k\mathbf{x}'_k\right)^{-1}\left(\sum_s a_{1k}a_{2k}\mathbf{x}_k y_k\right) \\ \hat{\mathbf{B}}_{(\mathbf{xB};\mathbf{x}_1)} &= \left(\sum_{s_1} a_{1k}\mathbf{x}_{1k}\mathbf{x}'_{1k}\right)^{-1}\left(\sum_{s_1} a_{1k}\mathbf{x}_{1k}\mathbf{x}'_k\hat{\mathbf{B}}_{(y;\mathbf{x})}\right)\end{aligned} \tag{5.6}$$

are the sample based estimates of $\mathbf{B}_{(y;\mathbf{x})}$ and $\mathbf{B}_{(\mathbf{xB};\mathbf{x}_1)}$ respectively. Note that we can calculate $\mathbf{x}'_k\hat{\mathbf{B}}_{(y;\mathbf{x})}$ in the second term of $\hat{\mathbf{B}}_{(\mathbf{xB};\mathbf{x}_1)}$ because $\mathbf{x}_k$ is known for $k \in s_1$.

Now let

$$R_1 = \left(\sum_U \mathbf{x}_{1k} - \sum_{s_1} a_{1k}\mathbf{x}_{1k}\right)' (\hat{\mathbf{B}}_{(\mathbf{xB};\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{xB};\mathbf{x}_1)})$$

and

$$R_2 = \left(\sum_{s_1} a_{1k}\mathbf{x}_k - \sum_s a_{1k}a_{2k}\mathbf{x}_k\right)' (\hat{\mathbf{B}}_{(y;\mathbf{x})} - \mathbf{B}_{(y;\mathbf{x})})$$

Both $N^{-1}R_1$ and $N^{-1}R_2$ are a product of two terms each converging to zero in probability under general conditions. In $N^{-1}R_1$, $N^{-1}(\sum_U \mathbf{x}_{1k} - \sum_{s_1} a_{1k}\mathbf{x}_{1k})$ is $O_p(n_1^{-1/2})$ and $(\hat{\mathbf{B}}_{(\mathbf{xB};\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{xB};\mathbf{x}_1)})$ is $O_p(n^{-1/2})$ since $\hat{\mathbf{B}}_{(\mathbf{xB};\mathbf{x}_1)}$ is a function of n through $\hat{\mathbf{B}}_{(y;\mathbf{x})}$ in $(\sum_{s_1} a_{1k}\mathbf{x}_{1k}\mathbf{x}'_k\hat{\mathbf{B}}_{(y;\mathbf{x})})$. Therefore $N^{-1}R_1$ is $O_p(n^{-1})$. Similarly in $N^{-1}R_2$,

$N^{-1}(\sum_{s_1} a_{1k}\mathbf{x}_k - \sum_s a_{1k}a_{2k}\mathbf{x}_k)$ is $O_p(n^{-1/2})$ and $(\hat{\mathbf{B}}_{(y;\mathbf{x})} - \mathbf{B}_{(y;\mathbf{x})})$ is $O_p(n^{-1/2})$ so $N^{-1}R_2$ is $O_p(n^{-1})$. The terms $N^{-1}R_1$ and $N^{-1}R_2$ are obviously of smaller order than $N^{-1}(\sum_{s_1} a_{1k}e_{k(\mathbf{xB};\mathbf{x}_1)})$ and $N^{-1}(\sum_s a_{1k}a_{2k}e_{k(y;\mathbf{x})})$ which are $O_p(n_1^{-1/2})$ and $O_p(n^{-1/2})$, respectively.

To obtain the bias of $\hat{Y}_{\mathrm{A}}$, we use (5.3) and express Y as

$$Y = \sum_U \mathbf{x}'_{1k}\mathbf{B}_{(\mathbf{xB};\mathbf{x}_1)} + \sum_U (e_{k(\mathbf{xB};\mathbf{x}_1)} + e_{k(y;\mathbf{x})}) \tag{5.7}$$

It follows from (5.5) that the bias is given by

$$\begin{aligned}\mathrm{Bias}(\hat{Y}_{\mathrm{A}}) = E\Bigg\{ &\left(\sum_U \mathbf{x}_{1k} - \sum_{s_1} a_{1k}\mathbf{x}_{1k}\right)' (\hat{\mathbf{B}}_{(\mathbf{xB};\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{xB};\mathbf{x}_1)}) \\ &+ \left(\sum_{s_1} a_{1k}\mathbf{x}_k - \sum_s a_{1k}a_{2k}\mathbf{x}_k\right)' (\hat{\mathbf{B}}_{(y;\mathbf{x})} - \mathbf{B}_{(y;\mathbf{x})})\Bigg\}\end{aligned} \tag{5.8}$$

The bias is simply $\mathrm{E}(R_1 + R_2)$. By the above analysis it is $O(n^{-1})$ and therefore close to zero.

Ignoring the lower order terms in (5.5), we obtain the approximation to $\hat{Y}_{\mathrm{A}}$ given by the linearized statistic

$$\hat{Y}_{\mathrm{A}} \cong \sum_U \mathbf{x}'_{1k}\mathbf{B}_{(\mathbf{xB};\mathbf{x}_1)} + \sum_{s_1} a_{1k}e_{k(\mathbf{xB};\mathbf{x}_1)} + \sum_s a_{1k}a_{2k}e_{k(y;\mathbf{x})} \tag{5.9}$$

The first term on the right-hand side is a constant. To obtain the approximate variance of $\hat{Y}_{\mathrm{A}}$, we condition on s_1 and apply the conditional variance rule to the right-hand side of (5.9), noting that $\mathrm{E}_{s|s_1}(\sum_s a_{1k}a_{2k}e_{k(y;\mathbf{x})}) = \sum_{s_1} a_{1k}e_{k(y;\mathbf{x})}$ and $e_{k(\mathbf{xB};\mathbf{x}_1)} + e_{k(y;\mathbf{x})} = y_k - \mathbf{x}'_{1k}\mathbf{B}_{(\mathbf{xB};\mathbf{x}_1)}$. This gives

$$\mathrm{V}(\hat{Y}_{\mathrm{A}}) \cong \mathrm{V}_{s_1}\left\{\sum_{s_1} a_{1k}(y_k - \mathbf{x}'_{1k}\mathbf{B}_{(\mathbf{xB};\mathbf{x}_1)})\right\} + \mathrm{E}_{s_1}\mathrm{V}_{s|s_1}\left\{\sum_s a_{1k}a_{2k}(y_k - \mathbf{x}'_k\mathbf{B}_{(y;\mathbf{x})})\right\} \tag{5.10}$$

Using the first three forms of $\mathbf{x}_k$ in (4.2), and Properties 1 and 2, this expression yields the approximate variance for A1, A2, and A3 in Table 2. For A4, we proceed as follows. To find the bias, we note that $\mathrm{E}_{s_1}\mathrm{E}_{s|s_1}(\sum_s w_{1k}a_{2k}y_k) = \mathrm{E}_{s_1}(\sum_{s_1} w_{1k}y_k)$ since the weights w_{1k} are independent of s. Then we replace y_k with its linear representation $y_k = \mathbf{x}'_{1k}\mathbf{B}_{(y;\mathbf{x}_1)} + e_{k(y;\mathbf{x}_1)}$ and use the calibration equation $\Sigma_{s_1} w_{1k}\mathbf{x}_{1k} = \Sigma_U \mathbf{x}_{1k}$. This leads to

$$\mathrm{Bias}(\hat{Y}_{\mathrm{A4}}) = E\left\{\left(\sum_U \mathbf{x}_{1k} - \sum_{s_1} a_{1k}\mathbf{x}_{1k}\right)' (\hat{\mathbf{B}}_{(y;\mathbf{x}_1)} - \mathbf{B}_{(y;\mathbf{x}_1)})\right\} \tag{5.11}$$

Similarly, we obtain the approximate variance of A4 as

$$\mathrm{V}(\hat{Y}_{\mathrm{A4}}) \cong \mathrm{V}_{s_1}\left\{\sum_{s_1} a_{1k}(y_k - \mathbf{x}'_{1k}\mathbf{B}_{(y;\mathbf{x}_1)})\right\} + \mathrm{E}_{s_1}\mathrm{V}_{s|s_1}\left\{\sum_s a_{1k}a_{2k}y_k\right\} \tag{5.12}$$

We include A4 in the first group because the calibration equation and the approximate variance for this case can be obtained by $\mathbf{x}_k = \phi$ in the general formulas for Case A.

Putting $\mathbf{x}_k = \phi$ eliminates the second calibration equation leaving the appropriate calibration equation for A4. Furthermore, with $\mathbf{x}_k = \phi$ in (5.10), we obtain the approximate variance for A4.

Similar analyses can be carried out for Cases B and C. For each of the ten estimators, we obtain an expression of the approximate variance given by

$$\mathrm{V}(\hat{Y}) \cong \mathrm{V}_{s_1}\left\{\sum_{s_1} a_{1k}e_{1k}\right\} + \mathrm{E}_{s_1}\mathrm{V}_{s|s_1}\left\{\sum_{s} a_{1k}a_{2k}e_{2k}\right\} \tag{5.13}$$

where e_{1k} and e_{2k} are given in Table 2. We make the following observations about the residuals e_{1k} and e_{2k}.

Remark 1. Suppose y is a linear combination of $\mathbf{x}_1$ and $\mathbf{x}_2$ given by $y_k = \mathbf{x}'_k\mathbf{B}_{(y;\mathbf{x})}$ for $k \in U$ with $\mathbf{x}'_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})$. It then follows that $e_{2k} = 0$ for A1, B1, and C1 and from (5.13) we see that these estimators reduce to one-phase estimators with approximate variance $\mathrm{V}_{s_1}\{\sum_{s_1} a_{1k}e_{1k}\}$. For A1, we can also interpret this result as follows. Suppose we have $\mathbf{x}_{1k}$ for $k \in s_1$ and the corresponding auxiliary total given by $\sum_U \mathbf{x}_{1k}$. If the vector of residuals $y_k - \mathbf{x}'_{1k}\mathbf{B}_{(\mathbf{x}_1;y)}$ lies in the column space of $\mathbf{x}_2$, it then follows that $e_{2k} = 0$. While this is unlikely to happen in any survey, it suggests that in order to produce an efficient estimator, we should look for an auxiliary $\mathbf{x}_2$ that explains as much of the variability of the residuals $y_k - \mathbf{x}'_{1k}\mathbf{B}_{(y;\mathbf{x}_1)}$ as possible.

Remark 2. When y is an exact linear combination of $\mathbf{x}_1$ written as $y_k = \mathbf{x}'_{1k}\mathbf{B}_{(y;\mathbf{x}_1)}$ for $k \in U$ then $e_{1k} = 0$ and $e_{2k} = 0$ for A1, A2, B1, and B2. This means that $\hat{Y}_{\mathrm{A1}} = \hat{Y}_{\mathrm{A2}} = \hat{Y}_{\mathrm{B1}} = \hat{Y}_{\mathrm{B2}} = Y$ for every two-phase sample. This result suggests that $\hat{Y}_{\mathrm{A1}}$, $\hat{Y}_{\mathrm{A2}}$, $\hat{Y}_{\mathrm{B1}}$ and $\hat{Y}_{\mathrm{B2}}$ are efficient estimators if the residuals $e_{k(y;\mathbf{x}_1)} = y_k - \mathbf{x}'_{1k}\mathbf{B}_{(y;\mathbf{x}_1)}$ are small. It is interesting to note that for A4, $e_{1k} = 0$ but $e_{2k} = y_k$ when $y_k = \mathbf{x}'_{1k}\mathbf{B}_{(y;\mathbf{x}_1)}$ for $k \in U$. Therefore, even in the unlikely situation of an exact regression of y on $\mathbf{x}_1$, the second

Case	e_1	e_2
A1	$e_{(y;\mathbf{x}_1)}$	$e_{(y;\mathbf{x})}$
A2	$e_{(y;\mathbf{x}_1)}$	$e_{(y;\mathbf{x}_1)}$
A3	$e_{(y;\mathbf{x}_2)} + e_{(\mathbf{x}_2\mathbf{B};\mathbf{x}_1)}$	$e_{(y;\mathbf{x}_2)}$
A4	$e_{(y;\mathbf{x}_1)}$	y
B1	$\mathbf{x}_2\mathbf{B}^{(2)}_{(y;\mathbf{x})} + e_{(y;\mathbf{x})}$	$e_{(y;\mathbf{x})}$
B2	$e_{(y;\mathbf{x}_1)}$	$e_{(y;\mathbf{x}_1)}$
C1	y	$e_{(y;\mathbf{x})}$
C2	y	$e_{(y;\mathbf{x}_1)}$
C3	y	$e_{(y;\mathbf{x}_2)}$
C4	y	y

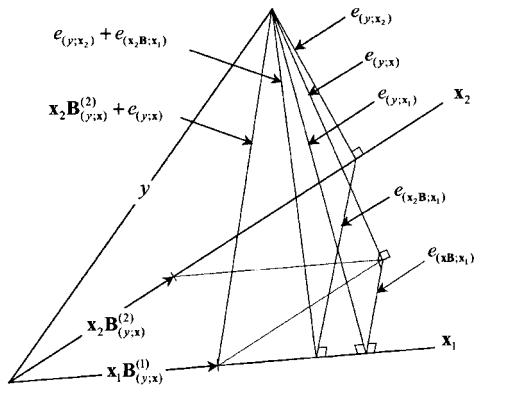

Fig. 1. Vector interpretation of residuals in the variance of two-phase calibration estimators

component of the variance of A4, which includes variability over both phases of selection, still remains.

The differences between the ten estimators are reflected in the two components of the approximate variance which are functions of e_{1k} and e_{2k}, respectively. Fig. 1 provides a geometric interpretation of these residuals. In this diagram, y, e_1, and e_2 are vectors in N-dimensional space. Furthermore, the lines shown for $\mathbf{x}_1$ and $\mathbf{x}_2$ should be interpreted as subspaces generated by the columns containing the auxiliary variable values over the population. The diagram provides insight into the general properties of the residuals and enables us to compare the variance of the different estimators. Fig. 1 also provides another way of expressing the residuals e_{1k} and e_{2k}. The vector $e_{(y;\mathbf{x})}$ is obtained from an OLS projection of y on $\mathbf{x}$, where $\mathbf{x}$ is the subspace generated by $\mathbf{x}_1$ and $\mathbf{x}_2$. Therefore the Euclidean magnitude of $e_{(y;\mathbf{x})}$ is less than or equal to that of $e_{(y;\mathbf{x}_1)}$, $e_{(y;\mathbf{x}_2)}$, $e_{(y;\mathbf{x}_2)} + e_{(\mathbf{x}_2\mathbf{B};\mathbf{x}_1)}$ and $\mathbf{x}_2'\mathbf{B}^{(2)}_{(y;\mathbf{x}_1,\mathbf{x}_2)} + e_{(y;\mathbf{x}_1,\mathbf{x}_2)}$. Similarly, the vectors $e_{(y;\mathbf{x}_1)}$ and $\mathbf{x}_2'\mathbf{B}^{(2)}_{(y;\mathbf{x}_1,\mathbf{x}_2)} + e_{(y;\mathbf{x}_1,\mathbf{x}_2)}$ are both projections of y on $\mathbf{x}_1$. However, by (5.1), $e_{(y;\mathbf{x}_1)}$ has minimum Euclidean distance because of the OLS projection of y on $\mathbf{x}_1$. Can we use these results to compare the approximate variance of the different estimators? For example, let us look at estimators A1 and B1. We note that they have the same approximate variance for the second component. Therefore, the difference in the approximate variance of these estimators is due to the variance of the residuals e_{1k} of the first component. Estimator A1 has residuals $e_{1k} = y_k - \mathbf{x}'_{1k}\mathbf{B}_{(y;\mathbf{x}_1)} = e_{k(y;\mathbf{x}_1)}$, which are on average smaller than the corresponding residuals $e_{1k} = y_k - \mathbf{x}'_{1k}\mathbf{B}^{(1)}_{(y;\mathbf{x}_1,\mathbf{x}_2)} = \mathbf{x}'_{2k}\mathbf{B}^{(2)}_{(y;\mathbf{x}_1,\mathbf{x}_2)} + e_{k(y;\mathbf{x}_1,\mathbf{x}_2)}$ for B1. Despite this result, we cannot conclude that the approximate variance of A1 is always smaller than that of B1. However, this is true for specific designs and choices of $\mathbf{x}_1$. For example, let us consider the design used in the simulations of Section 7 with SRS at each phase. It is easy to show that for this design, the approximate variance given by (5.13) simplifies to

$$\mathrm{V}(\hat{Y}) \cong N^2\left(1 - \frac{n_1}{N}\right)\frac{1}{n_1}\sum_U \frac{(e_{1k} - \bar{e}_1)^2}{N-1} + N^2\left(1 - \frac{n}{n_1}\right)\frac{1}{n}\sum_U \frac{(e_{2k} - \bar{e}_2)^2}{N-1} \tag{5.14}$$

where n_1 and n are the first-phase and second-phase sample sizes ($n \leqslant n_1$), $\bar{e}_1 = \sum_U e_{1k}/N$ and $\bar{e}_2 = \sum_U e_{2k}/N$. Since the population size N is known, we can calibrate on this total by including a count variable with value 1 in the auxiliary vector $\mathbf{x}_1$. It then follows from the normal equations that $\bar{e}_{(y;\mathbf{x}_1)} = \sum_U e_{k(y;\mathbf{x}_1)}/N = 0$ and $\bar{e}_{(y;\mathbf{x})} = \sum_U e_{k(y;\mathbf{x})}/N = 0$ with $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})'$. We can also derive the following results.

(1) $\sum_U (y_k - \bar{Y})^2 = \sum_U e^2_{k(y;\mathbf{x}_1)} + \sum_U (\mathbf{x}'_{1k}\mathbf{B}_{(y;\mathbf{x}_1)} - \bar{Y})^2$

(2) $\sum_U e^2_{k(y;\mathbf{x}_1)} = \sum_U e^2_{k(y;\mathbf{x})} + \sum_U (\mathbf{x}'_k\mathbf{B}_{(y;\mathbf{x})} - \mathbf{x}'_{1k}\mathbf{B}_{(y;\mathbf{x}_1)})^2$

(3) $\sum_U (e_{k(y;\mathbf{x}_2)} - \bar{e}_{(y;\mathbf{x}_2)})^2 = \sum_U e^2_{k(y;\mathbf{x})} + \sum_U (\mathbf{x}'_k\mathbf{B}_{(y;\mathbf{x})} - \mathbf{x}'_{2k}\mathbf{B}_{(y;\mathbf{x}_2)} - \bar{e}_{(y;\mathbf{x}_2)})^2$

(4) $\sum_U ((y_k - \mathbf{x}'_{1k}\mathbf{B}_{(\mathbf{x}_2\mathbf{B};\mathbf{x}_1)}) - \bar{e}_1^{(A3)})^2 = \sum_U e^2_{k(y;\mathbf{x}_1)} + \sum_U (\mathbf{x}'_{1k}(\mathbf{B}_{(y;\mathbf{x}_1)} - \mathbf{B}_{(\mathbf{x}_2\mathbf{B};\mathbf{x}_1)}) - \bar{e}_1^{(A3)})^2$
where $\bar{e}_1^{(A3)} = \sum_U (y_k - \mathbf{x}'_{1k}\mathbf{B}_{(\mathbf{x}_2\mathbf{B};\mathbf{x}_1)})/N$

(5) $\sum_U ((y_k - \mathbf{x}'_{1k}\mathbf{B}^{(1)}_{(y;\mathbf{x}_1,\mathbf{x}_2)}) - \bar{e}_1^{(B1)})^2 = \sum_U e^2_{k(y;\mathbf{x}_1)} + \sum_U (\mathbf{x}'_{1k}(\mathbf{B}_{(y;\mathbf{x}_1)} - \mathbf{B}^{(1)}_{(y;\mathbf{x}_1,\mathbf{x}_2)}) - \bar{e}_1^{(B1)})^2$
where $\bar{e}_1^{(B1)} = \sum_U (y_k - \mathbf{x}'_{1k}\mathbf{B}^{(1)}_{(y;\mathbf{x}_1,\mathbf{x}_2)})/N$

Results (1) to (5) allow us to compare the approximate variance of the ten estimators for this particular design and choice of auxiliary vector $\mathbf{x}_1$. For example, Result (5) implies $V(\hat{Y}_{\mathrm{A1}}) \leqslant V(\hat{Y}_{\mathrm{B1}})$. Comparisons can be made between other pairs of estimators. We show the result of these comparisons as a tree diagram in Fig. 2. A link between two

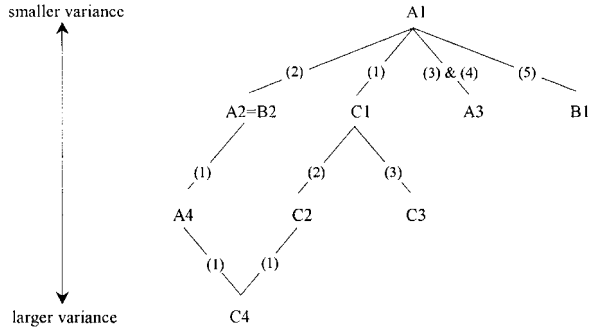

Fig. 2. Comparison of the approximate variance of the ten estimators for a design with SRS at each phase and with the count variable included in $\mathbf{x}_1$

estimators means that the higher one has variance smaller than or equal to that of the lower one. The number on the link shows the result required to prove each comparison. The approximate variance of A1 is smaller than or equal to that of any other estimator. Hence, it appears at the top of the tree. We emphasize that the comparisons given by Fig. 2 are not valid when $\mathbf{x}_1$ does not include the count variable. This can be seen from the results in Tables 6 and 7 of Section 7.

6. Variance Estimation

Variance estimates for the ten estimators can be obtained from the approximate variance given by (5.13). If we follow the standard approach to variance estimation, such as the one given by Särndal, Swensson, and Wretman (1992), we obtain a general form of the variance estimate as

$$\hat{\mathrm{V}}_1(\hat{Y}) = \sum_{k\in s}\sum_{l\in s}\frac{\Delta_{1kl}}{\pi_{1kl}\pi_{2kl}}\frac{\hat{e}_{1k}}{\pi_{1k}}\frac{\hat{e}_{1l}}{\pi_{1l}} + \sum_{k\in s}\sum_{l\in s}\frac{\Delta_{2kl}}{\pi_{2kl}}\frac{\hat{e}_{2k}}{\pi_{1k}\pi_{2k}}\frac{\hat{e}_{2l}}{\pi_{1l}\pi_{2l}} \tag{6.1}$$

In this expression, π_{1kl} is the first-phase joint inclusion probability of units k and l and π_{2kl} is the second-phase (conditional) joint inclusion probability of units k and l given $(k,l) \in s_1$. In addition we have $\Delta_{1kl} = \pi_{1kl} - \pi_{1k}\pi_{1l}$ and $\Delta_{2kl} = \pi_{2kl} - \pi_{2k}\pi_{2l}$. The residuals $\hat{e}_{1k}$ and $\hat{e}_{2k}$ are calculated by first estimating the unknown parameters in e_{1k} and e_{2k}, respectively. The residuals and parameter estimates are shown in Table 3.

Each term in (6.1) is an estimate of the corresponding term in variance formula (5.13). This approach works for all ten estimators but on closer inspection of (5.13) we note that we can do better for the estimated variance of $\hat{Y}_{\mathrm{A1}}$, $\hat{Y}_{\mathrm{A2}}$ and $\hat{Y}_{\mathrm{A3}}$. For these estimators, we can calculate the estimated regression coefficient $\hat{\mathbf{B}}_{(y;\mathbf{x})}$ over the units in s, use the predicted values $\mathbf{x}'_k\hat{\mathbf{B}}_{(y;\mathbf{x})}$ over s_1 to calculate the second regression coefficient $\hat{\mathbf{B}}_{(\mathbf{xB};\mathbf{x}_1)}$ and then obtain $\hat{e}_{1k}$ for $k \in s_1$. This allows us to obtain a better estimate of the first component of variance as a double sum over s_1 instead of a double sum over s. Thus, we obtain the estimated variance of $\hat{Y}_{\mathrm{A1}}$, $\hat{Y}_{\mathrm{A2}}$ and $\hat{Y}_{\mathrm{A3}}$ as

$$\hat{\mathrm{V}}_2(\hat{Y}) = \sum_{k\in s_1}\sum_{l\in s_1}\frac{\Delta_{1kl}}{\pi_{1kl}}\frac{\hat{e}_{1k}}{\pi_{1k}}\frac{\hat{e}_{1l}}{\pi_{1l}} + \sum_{k\in s}\sum_{l\in s}\frac{\Delta_{2kl}}{\pi_{2kl}}\frac{\hat{e}_{2k}}{\pi_{1k}\pi_{2k}}\frac{\hat{e}_{2l}}{\pi_{1l}\pi_{2l}} \tag{6.2}$$

Table 3. Residuals and parameter estimates for variance estimation

Case	Auxiliary $\mathbf{x}$	Residuals in estimated variance (6.1) and (6.2)		Parameter estimates
		$\hat{e}_{1k}$	$\hat{e}_{2k}$	
A1	$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_{1k} \\ \mathbf{x}_{2k} \end{pmatrix}$	$y_k - \mathbf{x}'_{1k}\hat{\mathbf{B}}_{(\mathbf{xB};\mathbf{x}_1)}$	$y_k - \mathbf{x}'_k\hat{\mathbf{B}}_{(y;\mathbf{x})}$	$\hat{\mathbf{B}}_{(y;\mathbf{x})} = \left(\sum_s a_{1k}a_{2k}\mathbf{x}_k\mathbf{x}'_k\right)^{-1}\sum_s a_{1k}a_{2k}\mathbf{x}_k y_k$ $\hat{\mathbf{B}}_{(\mathbf{xB};\mathbf{x}_1)} = \left(\sum_{s_1} a_{1k}\mathbf{x}_{1k}\mathbf{x}'_{1k}\right)^{-1}\sum_{s_1} a_{1k}\mathbf{x}_{1k}\mathbf{x}'_k\hat{\mathbf{B}}_{(y;\mathbf{x})}$
A2	$\mathbf{x}_k = \mathbf{x}_{1k}$	$y_k - \mathbf{x}'_{1k}\hat{\mathbf{B}}_{(y;\mathbf{x}_1)}$	$y_k - \mathbf{x}'_{1k}\hat{\mathbf{B}}_{(y;\mathbf{x}_1)}$	$\hat{\mathbf{B}}_{(y;\mathbf{x}_1)} = \left(\sum_s a_{1k}a_{2k}\mathbf{x}_{1k}\mathbf{x}'_{1k}\right)^{-1}\sum_s a_{1k}a_{2k}\mathbf{x}_{1k}y_k$
A3	$\mathbf{x}_k = \mathbf{x}_{2k}$	$y_k - \mathbf{x}'_{1k}\hat{\mathbf{B}}_{(\mathbf{x}_2\mathbf{B};\mathbf{x}_1)}$	$y_k - \mathbf{x}'_{2k}\hat{\mathbf{B}}_{(y;\mathbf{x}_2)}$	$\hat{\mathbf{B}}_{(y;\mathbf{x}_2)} = \left(\sum_s a_{1k}a_{2k}\mathbf{x}_{2k}\mathbf{x}'_{2k}\right)^{-1}\sum_s a_{1k}a_{2k}\mathbf{x}_{2k}y_k$ $\hat{\mathbf{B}}_{(\mathbf{x}_2\mathbf{B};\mathbf{x}_1)} = \left(\sum_{s_1} a_{1k}\mathbf{x}_{1k}\mathbf{x}'_{1k}\right)^{-1}\sum_{s_1} a_{1k}\mathbf{x}_{1k}\mathbf{x}'_{2k}\hat{\mathbf{B}}_{(y;\mathbf{x}_2)}$
A4	$\mathbf{x}_k = \phi$	$y_k - \mathbf{x}'_{1k}\hat{\mathbf{B}}_{(y;\mathbf{x}_1)}$	y_k	$\hat{\mathbf{B}}_{(y;\mathbf{x}_1)} = \left(\sum_s a_{1k}a_{2k}\mathbf{x}_{1k}\mathbf{x}'_{1k}\right)^{-1}\sum_s a_{1k}a_{2k}\mathbf{x}_{1k}y_k$
B1	$\mathbf{x}_k = \mathbf{x}_{2k}$	$y_k - \mathbf{x}'_{1k}\hat{\mathbf{B}}^{(1)}_{(y;\mathbf{x}_1,\mathbf{x}_2)}$	$y_k - (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})\hat{\mathbf{B}}_{(y;\mathbf{x}_1,\mathbf{x}_2)}$	$\hat{\mathbf{B}}_{(y;\mathbf{x}_1,\mathbf{x}_2)} = \left(\sum_s a_{1k}a_{2k}\begin{pmatrix}\mathbf{x}_{1k}\\ \mathbf{x}_{2k}\end{pmatrix}(\mathbf{x}'_{1k}, \mathbf{x}'_{2k})\right)^{-1}\sum_s a_{1k}a_{2k}\begin{pmatrix}\mathbf{x}_{1k}\\ \mathbf{x}_{2k}\end{pmatrix}y_k$
B2	$\mathbf{x}_k = \phi$	$y_k - \mathbf{x}'_{1k}\hat{\mathbf{B}}_{(y;\mathbf{x}_1)}$	$y_k - \mathbf{x}'_{1k}\hat{\mathbf{B}}_{(y;\mathbf{x}_1)}$	$\hat{\mathbf{B}}_{(y;\mathbf{x}_1)} = \left(\sum_s a_{1k}a_{2k}\mathbf{x}_{1k}\mathbf{x}'_{1k}\right)^{-1}\sum_s a_{1k}a_{2k}\mathbf{x}_{1k}y_k$
C1	$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_{1k} \\ \mathbf{x}_{2k} \end{pmatrix}$	y_k	$y_k - (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})\hat{\mathbf{B}}_{(y;\mathbf{x}_1,\mathbf{x}_2)}$	$\hat{\mathbf{B}}_{(y;\mathbf{x}_1,\mathbf{x}_2)} = \left(\sum_s a_{1k}a_{2k}\begin{pmatrix}\mathbf{x}_{1k}\\ \mathbf{x}_{2k}\end{pmatrix}(\mathbf{x}'_{1k}, \mathbf{x}'_{2k})\right)^{-1}\sum_s a_{1k}a_{2k}\begin{pmatrix}\mathbf{x}_{1k}\\ \mathbf{x}_{2k}\end{pmatrix}y_k$
C2	$\mathbf{x}_k = \mathbf{x}_{1k}$	y_k	$y_k - \mathbf{x}'_{1k}\hat{\mathbf{B}}_{(y;\mathbf{x}_1)}$	$\hat{\mathbf{B}}_{(y;\mathbf{x}_1)} = \left(\sum_s a_{1k}a_{2k}\mathbf{x}_{1k}\mathbf{x}'_{1k}\right)^{-1}\sum_s a_{1k}a_{2k}\mathbf{x}_{1k}y_k$
C3	$\mathbf{x}_k = \mathbf{x}_{2k}$	y_k	$y_k - \mathbf{x}'_{2k}\hat{\mathbf{B}}_{(y;\mathbf{x}_2)}$	$\hat{\mathbf{B}}_{(y,\mathbf{x}_2)} = \left(\sum_s a_{1k}a_{2k}\mathbf{x}_{2k}\mathbf{x}'_{2k}\right)^{-1}\sum_s a_{1k}a_{2k}\mathbf{x}_{2k}y_k$
C4	$\mathbf{x}_k = \phi$	y_k	y_k	(none)

This approach to variance estimation was noted by Axelson (2000). It makes better use of the auxiliary data for the estimation of the variance of $\hat{Y}_{A1}$, $\hat{Y}_{A2}$ and $\hat{Y}_{A3}$.

7. Simulation and Results

This section describes the design and the results of the simulation. To represent different conditions, we created six artificial populations, each of size $N = 1{,}000$. All six involve a variable y, generated as a linear function of one or both of the single auxiliary variables, x_1 and x_2, and a random error term. The theoretical variance of y is set to be the same for all six populations. This makes the double expansion estimator have roughly the same simulation variance in all populations and facilitates a comparison of results across the populations. The information used for calibration may reach the level of the whole population for x_1 but is limited to the level of the first-phase sample for x_2. The populations differ with respect to two factors: (i) the linear regression relationship of y on x_1 and x_2 (three types: a linear function of both x_1 and x_2, of x_1 only, of x_2 only); (ii) the size of the correlation between x_1 and x_2 (two types: near 0, near 1). This results in six populations for which a convenient notation is introduced in the body of the following 3×2 table.

Regression of y on x_1 and x_2	Correlation between x_1 and x_2	
	Near 0	Near 1
Linear in x_1 and x_2	$U_{12(0)}$	$U_{12(1)}$
Linear in x_1 only	$U_{1(0)}$	$U_{1(1)}$
Linear in x_2 only	$U_{2(0)}$	$U_{2(1)}$

Populations $U_{12(0)}$, $U_{1(0)}$ and $U_{2(0)}$ were constructed as follows: For each unit k, we generated independently $x_{1k} \sim \text{Gamma}(9, 10)$, $x_{2k} \sim \text{Gamma}(9, 10)$ and $\varepsilon_k \sim \text{Normal}(0, 25^2)$; $k = 1, \ldots, 1{,}000$. (The Gamma(a, b) distribution has density $f(x) = [\Gamma(a)\, b^a]^{-1} x^{a-1} \exp(-x/b)$ for $x > 0$, with $\text{E}(x) = a\, b$ and $\text{Var}(x) = a\, b^2$.) Then, for $k = 1, \ldots, 1{,}000$, the value of the variable of interest, y_k, was computed for each population as follows: $U_{12(0)}$: $y_k = x_{1k} + x_{2k} + \varepsilon_k$; $U_{1(0)}$: $y_k = \sqrt{2} x_{1k} + \varepsilon_k$; $U_{2(0)}$: $y_k = \sqrt{2} x_{2k} + \varepsilon_k$. For all three populations, the theoretical correlation between x_1 and x_2 is 0 and the theoretical y-variance is $2{,}425 = 2 \times 900 + 625$. Thus, we obtained three populations of triples (y_k, x_{1k}, x_{2k}), $k = 1, \ldots, 1{,}000$. Each has a correlation between x_1 and x_2 close to 0 and a y-variance near 2,425.

We include $U_{1(0)}$ and $U_{2(0)}$ in order to illustrate ''counterproductive calibration.'' For example, this can occur when the variable y is linearly related to only one of x_1 and x_2, but the calibration only uses information about the other variable. As shown by the simulation, this calibration can produce an estimator with larger variance than that of the double expansion estimator $\hat{Y}_{C4}$. This is important because we are often not in a position to determine whether y depends on one or both of the auxiliary variables x_1 and x_2.

Populations $U_{12(1)}$, $U_{1(1)}$ and $U_{2(1)}$ were constructed as follows. For unit k, we generated independently $x_{1k} \sim \text{Gamma}(9, 10)$ and $\varepsilon_k \sim \text{Normal}(0, 25^2)$; then we computed $x_{2k} = x_{1k} + \delta_k$, $k = 1, \ldots, 1000$, where $\delta_k \sim N(0, 45)$, independently of x_{1k} and ε_k. As a result of the comparatively small variance of δ_k, the correlation between x_1 and x_2

is near 1; its theoretical value is $\sqrt{900/945} = 0.976$. Then, for $k = 1, \ldots, 1000$, y_k was computed as follows: $U_{12(1)}$: $y_k = (2\sqrt{10}/9)(x_{1k} + x_{2k}) + \varepsilon_k$; $U_{1(1)}$: $y_k = \sqrt{2}x_{1k} + \varepsilon_k$; $U_{2(1)}$: $y_k = (2\sqrt{210}/21)x_{2k} + \varepsilon_k$. For all three, the theoretical y-variance is again 2,425. This gives three additional finite populations (y_k, x_{1k}, x_{2k}), $k = 1, \ldots, 1000$, all with a correlation between x_1 and x_2 close to 0.98 and a y-variance near 2,425.

We selected 100,000 independent two-phase samples from each of the six populations. Each first-phase sample s_1 was drawn as an SRS of size $n_1 = 500$ from the population U of size $N = 1,000$; for each s_1, s was drawn as an SRS of size $n = 200$ from s_1. Calibrated weights w_k were obtained for each (s_1, s) realization. They were determined for each of the ten cases and two different options for the auxiliary vectors: (i) $\mathbf{x}_{1k} = (1, x_{1k})'$; $\mathbf{x}_{2k} = x_{2k}$, and (ii) $\mathbf{x}_{1k} = x_{1k}$; $\mathbf{x}_{2k} = x_{2k}$.

Calibration using option (i) requires that the population size N be known, in addition to $\sum_U x_{1k}$. In our simulation, N is known, so from the standpoint of using all available auxiliary information, the natural choice is $\mathbf{x}_{1k} = (1, x_{1k})'$ rather than $\mathbf{x}_{1k} = x_{1k}$. However, some statisticians would prefer $\mathbf{x}_{1k} = x_{1k}$, arguing that if the regression of y on x_1 goes through the origin, allowing for an intercept is inappropriate.

For each of the 100,000 realizations of (s_1, s), calibration estimates were obtained for the ten estimators under the two options given above. A Monte Carlo bias was computed as the mean of the 100,000 estimates minus Y. Since the bias was negligible for all estimators, we omitted it from the simulation tables and analysis. This allowed us to focus on the variance rather than the mean squared error of each estimator. A simulation variance was calculated as the Monte Carlo variance of the 100,000 calibration estimates. This variance is shown as SimVar in Tables 4, 5, 6 and 7. The terms ApproxVE and ApproxEV represent the components of the approximate variance under SRS at each phase. These were calculated from the corresponding terms of formula (5.14) using the expressions in Table 2. This can be done since we can compute the residuals e_{1k} and e_{2k} for all $k \in U$. The approximate variance denoted by ApproxVar is simply the sum of ApproxVE and ApproxEV. The simulation results are summarized in Tables 4 and 5 for option (i) and in Tables 6 and 7 for option (ii). In all tables and for all cases, SimVar and ApproxVar are very close. This confirms formula (5.14) and the theory behind the linear representations and residuals given in Table 2.

We look for answers to the following three questions in the simulation results of Tables 4 to 7:

1. Should A1 have the smallest variance because it is the only one to use all the auxiliary information?
2. Should B1 and A1 have nearly identical variances, because they differ only slightly in the auxiliary information used for calibration?
3. Should C4 have the largest variance because it is the only case that uses no auxiliary information?

Tables 4 to 7 show that the answers to these questions are not always what we expect. We look into the reasons for this. In the following comments, ''improves on'' or ''is better than'' is to be understood as ''has smaller variance than.'' The term ''best'' means ''has the smallest variance'' among a set of alternatives. The term ''significantly better''

Table 4. Simulation results with $x_{1k} \sim Gamma(9, 10)$, $x_{2k} \sim Gamma(9, 10)$, $\epsilon_k \sim Normal(0, 25^2)$ and calibration variables $\mathbf{x}_{1k} = (1, x_{1k})'$ and $\mathbf{x}_{2k} = x_{2k}$. Variances are the displayed values $\times 10^6$

Population	Population model	Estimator	SimVar	ApproxVar	ApproxVE	ApproxEV
$U_{12(0)}$ $Y = 181113.90$	$y_k = x_{1k} + x_{2k} + \epsilon_k$	A1	3.29	3.25	1.46	1.79
		A2	5.91	5.85	1.46	4.39
		A3	9.17	9.12	2.37	6.75
		A4	8.65	8.47	1.46	7.01
		B1	3.29	3.25	1.46	1.79
		B2	5.91	5.85	1.46	4.39
		C1	4.15	4.13	2.34	1.79
		C2	6.75	6.73	2.34	4.39
		C3	9.13	9.09	2.34	6.75
		C4	9.38	9.35	2.34	7.01
$U_{1(0)}$ $Y = 128138.98$	$y_k = \sqrt{2}x_{1k} + \epsilon_k$	A1	2.43	2.39	0.60	1.79
		A2	2.42	2.40	0.60	1.80
		A3	14.21	14.12	2.41	11.71
		A4	7.89	7.74	0.60	7.14
		B1	2.43	2.39	0.60	1.79
		B2	2.42	2.40	0.60	1.80
		C1	4.20	4.17	2.38	1.79
		C2	4.19	4.18	2.38	1.80
		C3	14.18	14.09	2.38	11.71
		C4	9.61	9.52	2.38	7.14
$U_{2(0)}$ $Y = 127410.26$	$y_k = \sqrt{2}x_{2k} + \epsilon_k$	A1	4.17	4.14	2.35	1.79
		A2	9.47	9.41	2.35	7.06
		A3	4.15	4.14	2.35	1.79
		A4	9.47	9.41	2.35	7.06
		B1	4.17	4.14	2.35	1.79
		B2	9.47	9.41	2.35	7.06
		C1	4.16	4.14	2.35	1.79
		C2	9.46	9.41	2.35	7.06
		C3	4.14	4.14	2.35	1.79
		C4	9.41	9.41	2.35	7.06

or ''significantly worse'' is used when there is a difference of 5% or more between the variances, and ''insignificant'' refers to a difference of less than 5%.

By examining SimVar in Tables 4 and 5, we are led to the following observations for option (i) with $\mathbf{x}_{1k} = (1, x_{1k})'$ and $\mathbf{x}_{2k} = x_{2k}$:

1. As we can see in Table 5, A1 has smaller variance than the next best estimator only for $U_{12(1)}$. For all six populations, one or more estimators come very close to A1. For example, there are no significant differences between A1, A2, B1, and B2 for $U_{1(1)}$.
2. A1 is not always distinctly better than B1. For $U_{12(0)}$, $U_{1(0)}$ and $U_{2(0)}$ in Table 4, A1 and B1 are about the same, and they are close for $U_{12(1)}$ and $U_{1(1)}$ in Table 5. But $U_{2(1)}$ stands out, in that B1 has a much (about 50%) larger variance than A1.
3. For $U_{12(1)}$, $U_{1(1)}$, and $U_{2(1)}$ in Table 5, C4 has distinctly larger variance than all other cases. For $U_{12(0)}$, $U_{1(0)}$, and $U_{2(0)}$ in Table 4, there are other estimators with about the same variance as C4. For $U_{1(0)}$, we see some striking examples of counterproductive calibration, in that A3 and C3 have much larger variance than C4. With these two estimators, we unknowingly make a mistake by resorting to this calibration. It would be better to ignore the auxiliary information in these two cases.

Table 5. Simulation results with $x_{1k} \sim Gamma(9, 10)$, $x_{2k} = x_{1k} + \delta_k$ *with* $\delta_k \sim Normal(0, 45)$, $\epsilon_k \sim Normal(0, 25^2)$ *and calibration variables* $\mathbf{x}_{1k} = (1, x_{1k})'$ *and* $\mathbf{x}_{2k} = x_{2k}$. *Variances are the displayed values* $\times 10^6$

Population	Population model	Estimator	SimVar	ApproxVar	ApproxVE	ApproxEV
$U_{12(1)}$ $Y = 127419.33$	$y_k = (2\sqrt{10}/9)(x_{1k} + x_{2k}) + \epsilon_k$	A1	2.43	2.40	0.61	1.79
		A2	2.45	2.43	0.61	1.82
		A3	2.52	2.51	0.61	1.90
		A4	7.85	7.71	0.61	7.10
		B1	2.68	2.60	0.81	1.79
		B2	2.45	2.43	0.61	1.82
		C1	4.18	4.16	2.37	1.79
		C2	4.19	4.19	2.37	1.82
		C3	4.27	4.27	2.37	1.90
		C4	9.54	9.47	2.37	7.10
$U_{1(1)}$ $Y = 128138.98$	$y_k = \sqrt{2}x_{1k} + \epsilon_k$	A1	2.43	2.39	0.60	1.79
		A2	2.42	2.40	0.60	1.80
		A3	2.76	2.74	0.60	2.14
		A4	7.89	7.74	0.60	7.14
		B1	2.52	2.44	0.65	1.79
		B2	2.42	2.40	0.60	1.80
		C1	4.19	4.18	2.38	1.80
		C2	4.19	4.18	2.38	1.80
		C3	4.53	4.52	2.38	2.14
		C4	9.61	9.52	2.38	7.14
$U_{2(1)}$ $Y = 125183.84$	$y_k = (2\sqrt{210}/21)x_{2k} + \epsilon_k$	A1	2.48	2.45	0.66	1.79
		A2	2.65	2.63	0.66	1.97
		A3	2.47	2.45	0.66	1.79
		A4	7.85	7.72	0.66	7.06
		B1	3.71	3.64	1.85	1.79
		B2	2.65	2.63	0.66	1.97
		C1	4.16	4.14	2.35	1.79
		C2	4.32	4.32	2.35	1.97
		C3	4.15	4.14	2.35	1.79
		C4	9.48	9.41	2.35	7.06

We find more counterintuitive results when we examine SimVar in Tables 6 and 7 for option (ii) with $\mathbf{x}_{1k} = x_{1k}$ and $\mathbf{x}_{2k} = x_{2k}$:

1. A1 is not distinctly better than any other estimator for any population. For $U_{12(0)}$ and $U_{2(0)}$ in Table 6, B1 provides a significant improvement on A1 (by about 20%).
2. We find a mixed pattern when comparing A1 and B1 in the two tables. For $U_{12(0)}$ and $U_{2(0)}$, B1 is better than A1. By contrast, B1 has about 50% larger variance than A1 for $U_{2(1)}$. For $U_{1(0)}$, $U_{12(1)}$, and $U_{1(1)}$, B1 and A1 show no significant difference.
3. For $U_{12(1)}$, $U_{1(1)}$, and $U_{2(1)}$ in Table 7, C4 has distinctly larger variance than all other cases. For $U_{12(0)}$, A4 and C3 come close in variance to C4. In Table 6, we find several counterproductive calibrations for $U_{1(0)}$ and $U_{2(0)}$. For $U_{1(0)}$, A3 and C3 have much larger variance (about 50% for C3) than C4. Even more striking, A2, A4, B2, and C2 have much larger variance than C4 for $U_{2(0)}$. In fact, A2 has about 70% larger variance than C4.

Our simulations have produced several examples where calibration based on all the available auxiliary information can be counterproductive. For example, we have noted cases where B1 has smaller variance than A1 although A1 uses more auxiliary information.

Table 6. Simulation results with $x_{1k} \sim Gamma(9, 10)$, $x_{2k} \sim Gamma(9, 10)$, $\epsilon_k \sim Normal(0, 25^2)$ and calibration variables $\mathbf{x}_{1k} = x_{1k}$ and $\mathbf{x}_{2k} = x_{2k}$. Variances are the displayed values $\times 10^6$

Population	Population model	Estimator	SimVar	ApproxVar	ApproxVE	ApproxEV
$U_{12(0)}$ $Y = 181113.90$	$y_k = x_{1k} + x_{2k} + \epsilon_k$	A1	4.00	3.99	2.20	1.79
		A2	8.82	8.79	2.20	6.59
		A3	8.68	8.67	1.92	6.75
		A4	9.26	9.21	2.20	7.01
		B1	3.28	3.26	1.47	1.79
		B2	8.82	8.79	2.20	6.59
		C1	4.14	4.13	2.34	1.79
		C2	8.91	8.93	2.34	6.59
		C3	9.13	9.09	2.34	6.75
		C4	9.38	9.35	2.34	7.01
$U_{1(0)}$ $Y = 128138.98$	$y_k = \sqrt{2}x_{1k} + \epsilon_k$	A1	2.42	2.39	0.60	1.79
		A2	2.41	2.40	0.60	1.80
		A3	12.43	12.37	0.66	11.71
		A4	7.82	7.74	0.60	7.14
		B1	2.42	2.39	0.60	1.79
		B2	2.41	2.40	0.60	1.80
		C1	4.19	4.17	2.38	1.79
		C2	4.18	4.18	2.38	1.80
		C3	14.18	14.09	2.38	11.71
		C4	9.61	9.52	2.38	7.14
$U_{2(0)}$ $Y = 127410.26$	$y_k = \sqrt{2}x_{2k} + \epsilon_k$	A1	5.62	5.61	3.82	1.79
		A2	15.34	15.28	3.82	11.46
		A3	5.60	5.61	3.82	1.79
		A4	10.90	10.88	3.82	7.06
		B1	4.16	4.14	2.35	1.79
		B2	15.34	15.28	3.82	11.46
		C1	4.16	4.14	2.35	1.79
		C2	13.83	13.81	2.35	11.46
		C3	4.15	4.14	2.35	1.79
		C4	9.41	9.41	2.35	7.06

We have also found examples where C4, which uses no auxiliary data, has smaller variance than other estimators that use auxiliary information. The reasons become evident on closer inspection. For example, the calibrations in A2, A4, and C2 rely exclusively on information about x_1. But for population $U_{2(0)}$, x_1 is uncorrelated with x_2 which is the sole predictor of y. Instead of improving the weights, the calibration on x_1 produces inefficient weights, as shown by the results in Table 6. We conclude the discussion on the simulations with a few additional comments.

(i) *The efficiency gains from calibration*. For $U_{1(0)}$, $U_{12(1)}$, $U_{1(1)}$, and $U_{2(1)}$, and both formulations of the vector $\mathbf{x}_1$, the variance is reduced from about 9.5×10^6 (C4) to about 2.4×10^6 (the best estimator for each population), a reduction of about 70%. For $U_{12(0)}$ and $U_{2(0)}$, the gains are about 55%. In other words, the best calibration estimator always provides a significant improvement over no calibration.

(ii) *The effect of an intercept term in the vector* $\mathbf{x}_1$. The use of $\mathbf{x}_{1k} = (1, x_{1k})'$ and $\mathbf{x}_{2k} = x_{2k}$ rather than $\mathbf{x}_{1k} = x_{1k}$ and $\mathbf{x}_{2k} = x_{2k}$ does not improve the Case A

Table 7. Simulation results with $x_{1k} \sim Gamma(9, 10)$, $x_{2k} = x_{1k} + \delta_k$ *where* $\delta_k \sim N(0, 45)$, $\epsilon_k \sim Normal(0, 25^2)$ *and calibration variables* $\mathbf{x}_{1k} = x_{1k}$ *and* $\mathbf{x}_{2k} = x_{2k}$. *Variances are the displayed values* $\times 10^6$

Population	Population model	Estimator	SimVar	ApproxVar	ApproxVE	ApproxEV
$U_{12(1)}$ $Y = 127419.33$	$y_k = (2\sqrt{10}/9)(x_{1k} + x_{2k}) + \epsilon_k$	A1	2.42	2.40	0.61	1.79
		A2	2.44	2.43	0.61	1.82
		A3	2.52	2.51	0.61	1.90
		A4	7.78	7.71	0.61	7.10
		B1	2.67	2.60	0.81	1.79
		B2	2.44	2.43	0.61	1.82
		C1	4.17	4.16	2.37	1.79
		C2	4.18	4.19	2.37	1.82
		C3	4.27	4.27	2.37	1.90
		C4	9.54	9.47	2.37	7.10
$U_{1(1)}$ $Y = 128138.98$	$y_k = \sqrt{2}x_{1k} + \epsilon_k$	A1	2.41	2.39	0.60	1.79
		A2	2.41	2.40	0.60	1.80
		A3	2.76	2.74	0.60	2.14
		A4	7.82	7.74	0.60	7.14
		B1	2.50	2.43	0.64	1.79
		B2	2.41	2.40	0.60	1.80
		C1	4.18	4.17	2.38	1.79
		C2	4.18	4.18	2.38	1.80
		C3	4.53	4.52	2.38	2.14
		C4	9.61	9.52	2.38	7.14
$U_{2(1)}$ $Y = 125183.84$	$y_k = (2\sqrt{210}/21)x_{2k} + \epsilon_k$	A1	2.47	2.45	0.66	1.79
		A2	2.63	2.63	0.66	1.97
		A3	2.47	2.45	0.66	1.79
		A4	7.79	7.72	0.66	7.06
		B1	3.70	3.65	1.86	1.79
		B2	2.63	2.63	0.66	1.97
		C1	4.15	4.14	2.35	1.79
		C2	4.31	4.32	2.35	1.97
		C3	4.15	4.14	2.35	1.79
		C4	9.48	9.41	2.35	7.06

estimators when x_1 and x_2 are highly correlated ($U_{12(1)}$, $U_{1(1)}$ and $U_{2(1)}$). However, for A1 the improvement is significant when these variables are essentially uncorrelated and y is not a linear function of only x_1 ($U_{12(0)}$ and $U_{2(0)}$).

(iii) *Recalibrating on the vector* $\mathbf{x}_1$. The difference between A1 and A3 is that in A1, we calibrate on $\mathbf{x}_1$ at each step, but in A3, we only calibrate on $\mathbf{x}_1$ at the first step. Does the repeated use of $\mathbf{x}_1$ in the calibration lead to a significant improvement? We find that considerable improvement does occur for $U_{12(0)}$ and $U_{1(0)}$, when the variables x_1 and x_2 are nearly uncorrected and y is not a function of x_2 alone.

8. Recommendations and Discussion

In practice, it is important to make the best use of the available auxiliary information so as to obtain the most efficient estimator possible. Ideally, we would like the set of auxiliary variables in $\mathbf{x}_1$ to be closely linearly related to y. Whether this holds or not, we can try to use any additional set of auxiliary variables in $\mathbf{x}_2$ to explain as much as possible the variation remaining in the residuals $y_k - \mathbf{x}'_{1k}\mathbf{B}_{(y;\mathbf{x}_1)}$.

We have seen that the approximate variance is the sum of two components, each of which is a function of population residuals. The variance depends not only on the size

of these residuals but also on the two sampling designs (one for each phase) and on the respective sample sizes, n_1 and n.

A quick and simple approach requiring no analysis is to calibrate on all of the available auxiliary information (which may be less extensive than in the complete information Case A1). As we have seen in the simulations, this may not always produce the best estimator. Occasionally, it may even result in an inefficient estimator, with a variance larger than that of the double expansion estimator. However, it is a reasonable approach to use in a routine production of estimates for many variables of interest.

We could go further and conduct an analysis for each variable of interest, computing the estimated variance of the different estimators from the given auxiliary information. We have shown in Section 6 how the variance is estimated for a general two-phase design. To single out the best (minimum variance) use of the available auxiliary information for every y-variable in a large survey is not easy, at least not without special effort and time consuming analysis. Still, this procedure is only a guide because the variance estimator is a random quantity, and sometimes, there may be little to choose between several of the available alternative uses of the auxiliary information.

9. References

Alexander, C.H. (1987). A Class of Methods for Using Person Controls in Household Weighting. Survey Methodology, 13, 183–198.

Armstrong, J. and St-Jean, H. (1994). Generalized Regression Estimation for a Two-Phase Sample of Tax Records. Survey Methodology, 20, 97–106.

Axelson, M. (2000). On Variance Estimation for the Two-phase Regression Estimator. Ph.D. thesis, Uppsala University, Sweden.

Binder, D.A. (1996). Linearization Methods for Single Phase and Two-Phase Samples: A Cookbook Approach. Survey Methodology, 22, 17–22.

Bankier, M.A. (1989). Generalized Least Squares Estimation under Poststratification. Proceedings of the American Statistical Association, Section on Survey Research Methods, 730–755.

Deville, J.C. and Särndal, C.E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376–382.

Deville, J.C., Särndal, C.E., and Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. Journal of the American Statistical Association, 88, 1013–1020.

Dupont, F. (1995). Alternative Adjustments When There Are Several Levels of Auxiliary Information. Survey Methodology, 21, 125–136.

Estevao, V.M. (1994). Calculation of G-Weights under Calibration and Bound Constraints. Report, Statistics Canada.

Estevao, V.M. and Särndal, C.E. (2000). A Functional Form Approach to Calibration. Journal of Official Statistics, 16, 379–399.

Hidiroglou, M.A. and Särndal, C.E. (1998). Use of Auxiliary Information for Two-Phase Sampling. Survey Methodology, 24, 11–20.

Huang, E. and Fuller, W.A. (1978). Non-Negative Regression Estimation in Sample Survey Data. Proceedings of the American Statistical Association, Section of Social Statistics, 300–305.

Lundström, S. (1997). Calibration as a Standard Method for Treatment of Nonresponse. Ph.D. thesis, Stockholm University, Sweden.

Särndal, C.E. and Swensson, B. (1987). A General View of Estimation for Two Phases of Selection with Applications to Two-Phase Sampling and Nonresponse. International Statistical Review, 55, 279–294.

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

Singh, A.C. and Mohl, C.A. (1996). Understanding Calibration Estimators in Survey Sampling. Survey Methodology, 22, 107–115.

Received January 2001
Revised January 2002