Article

Sociological Methods & Research
00(0) 1-25
© The Author(s) 2013
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0049124113507906
smr.sagepub.com

A Novel Calibration Estimator in Social Surveys

Antonio Arcos[1], José M. Contreras[1] and María M. Rueda[1]

Abstract

Social surveys generally assume that a sample of units (students, individuals, employees, . . .) is observed by two-stage selection from a finite population, which is grouped into clusters (schools, household, companies, . . .). This design involves sampling from two different populations: the population of schools or primary stage units and the population of students or second-stage units. Calibration estimators for student statistics can be defined by using combined information based on school totals and student totals. Auxiliary information from the units at the two stages can be calibrated by integrated weighting, as proposed by Lemaître and Dufour or Estevao and Särndal. Two calibration estimators for the population total based on unit weights are defined. The first estimator satisfies a calibration equation at the unit level, and the second one, at the cluster level. The proposed estimator shrinks the unit estimator toward the cluster. A simulation study based on two real populations is carried out to study the empirical performance of this shrinkage estimator. The populations studied were obtained from the Programme for International Student Assessment database and from the Spanish Household Budget Survey.

[1]University of Granada, Granada, Spain

Corresponding Author:
Antonio Arcos, Campus de Cartuja s/n, University of Granada, Granada 17071, Spain.
Email: arcos@ugr.es

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

Keywords

auxiliary information, social surveys, calibration estimators

Introduction

The aim of any social survey program is to provide information on several variables. A school or household survey is a particular type of social survey, in which we are interested in the characteristics of all or some members of the school or household. These characteristics typically include a subset of variables such as health, education, income, expenditure, employment status, use of various types of services, school grade and, in general, information on schools and students' academic performance.

In educational surveys, examples of these variables are the condition of public and nonpublic (private) education, school choice, school characteristics (such as type, lowest and highest grades obtained there, religious affiliation), student experiences, teacher feedback on the children's school performance and behavior, family involvement, factors affecting education, community support, and so on. This information has two levels of hierarchy: school and student. It is interesting to measure both school-level and student-level results from achievement tests, and thus identify factors affecting educational progress.

This is done in several programs produced at the U.S. National Center for Education Statistics, such as the National Assessment of Educational Progress (NAEP), which compiles and reports information on the academic performance of U.S. students, and International Indicators of Education Systems, which assesses educational experiences and trends in countries other than the United States. This is the case of the Organization for Economic Cooperation and Development (OECD) Programme for International Student Assessment (PISA). With respect to NAEP methodology, articles by Rust and Johnson (1992) and Johnson and Rust (1992) to Jia et al. (2010) reflect the importance of survey sampling in educational surveys.

Collecting data related to education follows a standard general pattern. The selected samples are based on a two-stage sample design: selection of schools (within strata) and selection of students within schools. Other types of survey, such as the National Household Education Surveys Program, have a similar design, since the population of clusters or primary stage units are households.

Sampling weights are needed in order to make valid inferences about the populations from the units were drawn and are computed to obtain unbiased

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

estimates of population characteristics. Moreover, this procedure makes it possible to achieve unbiased estimates of standard errors, see, for example, Pfeffermann et al. (1998) and Pfeffermann (1993).

In this context, auxiliary information is defined as a set of variables measured in the survey and for which information on the population distribution is available. This auxiliary information can be used to compute adjustment weights, and then population characteristics can be estimated using weighted rather than unweighted values.

Several weighting techniques can be applied using the available auxiliary information of a qualitative or a quantitative variable. Deville and Särndall (1992) and Deville, Särndal, and Sautory (1993) created a general framework for weighting, termed the calibration approach. Calibration provides a simple and practical approach for incorporating auxiliary information into the estimation. For example, in our context, auxiliary information from schools means that the total number of public or nonpublic schools is known before the sample of schools is drawn. The total number of male and female students is also known before the sample is taken. Note that variables other than qualitative ones can be used as auxiliary information. In the case of schools, auxiliary information can also refer to the number of students, the average size of English classes, the number of teachers, the school's budget allocation, and so on. With respect to students, auxiliary information can also refer to their age, number of brothers and sisters, and so on.

The increasing power and availability of computers greatly facilitates data collection (auxiliary variables). Estimation schemes such as calibration may reduce the variance of the estimates by making use of auxiliary variables. Calibration estimators have various desirable properties: First, the weights provide sample "estimates" for the totals of the auxiliary variables that match known population totals for these variables. If the population totals of the auxiliary variables have been published before the survey results are produced, then using calibration estimators for the survey would guarantee that the survey estimates are coherent with those already in the public domain. The second desirable property is that of simplicity, the fact that given the weights, calibration estimates are linear in the main variable. The third desirable property of calibration estimators is their flexibility to incorporate auxiliary information, including continuous, discrete, or both types of benchmark variables at the same time. Calibration weighting has the advantage that it is relatively straightforward to compute variances of weighted estimators. However, there are no simple formulae for the variance of multiplicative weighting estimates.

Calibration adjustments may be used to accomplish several goals simultaneously, such as reducing sampling error, reducing residual nonresponse

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

bias and noncoverage bias, and matching known control totals. In the past decade, numerous articles have been produced with this goal in mind.

Several national statistical agencies have developed software designed to compute weights, usually calibrated in accordance with the auxiliary information available in administrative records and other accurate sources. The software CALMAR (Deville, Särndal, and Sautory 1993), CALMAR2, (LeGuennec and Sautory 2002), GES (Statistics Canada), CLAN97 (Statistics Sweden), Bascula 4.0 (Central Bureau of Statistics, the Netherlands), gCALIBS (Statistics Belgium), or the Sampling R-package (Tillé and Matei 2009) are noteworthy examples.

Calibration weights depend on the available auxiliary data. The data available from household surveys are quite different from those obtained in business or schools surveys. Auxiliary information may be available for both the clusters (households, schools, and companies) and the units, which creates a certain complexity regarding auxiliary information.

One means of addressing this complexity, and simultaneously of obtaining elementary estimates for units (individuals) and clusters (households or schools), through an integrated weighting technique, can be seen in Lemaître and Dufour (1987). Recent references are Gambino and Silva (2009) for household or school surveys and Särndal (2007) for calibration estimation.

The second section describes calibration estimation under two-stage sampling, together with different methods to incorporate complex auxiliary information. In the third and fourth sections, we propose a shrinkage method for weighting calibration estimators based on individuals (elementary units) and clusters (households or schools). The term shrinkage has multiple meanings depending on the context. Historically, the concept derives from James and Stein (1961), who proposed a nonlinear estimator which is more efficient than the ordinary least squared estimator. In statistics, shrinkage represents a technique to enhance an estimator, by improving the efficiency of the resulting estimator with respect to the mean squared error (MSE) criterion. A simulation study is described the fifth section. And, finally, the main conclusions are drawn in the sixth section.

General Setup

The usual formulation of sampling in two stages is as follows. A sample is drawn under a two-stage selection from a finite population U, which is grouped into clusters. This design involves sampling from two distinct populations: the population of clusters or primary stage units, $U_I = \{1, \ldots, i, \ldots, N_I\}$, and the

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

population of units or second-stage units, $U = \{1, \ldots, k, \ldots, N\}$, which is the union of all units in all N_I clusters U_i, $i \in U_I$.

First, we draw from U_I a sample of clusters S_I with inclusion probabilities π_{Ii}, with $i \in U_I$. The first-stage design weights are $d_{Ii} = 1/\pi_{Ii}$ for $i \in U_I$. At the second stage, units are selected from each of the selected clusters. For $i \in s_I$, we enumerate the units in U_i and draw a sample s_i from these units, with conditional inclusion probabilities $\pi_{k|i}$ and $k \in s_i$. The second-stage design weights are $d_{k|i}= 1/\pi_{k|i}$ for $k \in s_i$. Then $d_k=d_{Ii}d_{k|i}$ is the overall design weight for the kth unit, and $s = \cup_{i\in s_I} s_i$ is the sample of units.

In addition, data are often collected at each level of the hierarchy. For example, in the case of school surveys, data are collected at the school level and at the student level. It is then important to recognize and note which variables are measured at the student level and which are measured at the school level (e.g., socioeconomic characteristics of the school). Sampling is focused not only on households, where cluster sizes vary little, but also on situations in which cluster sizes may vary widely, such as educational survey, where in the sampling process, students are grouped by schools, large and small, or industrial surveys, where employees are grouped within firms (which may also have widely varying numbers of members/employees).

Many two-stage survey designs have a twofold goal: to estimate totals for the population of units U and for that of clusters U_I. Auxiliary information may be available for both clusters and units. This creates a certain complexity in the auxiliary information, and thus several methods can be used at the estimation stage. Let $y_{(u)}$ be a variable of interest defined for the units, where $y_{(u)k}$ is the value of $y_{(u)}$ for the unit k, with $k \in s$. Let $y_{(c)i}$ be the value of the variable of interest $y_{(c)}$ for the ith cluster, with $i \in s_I$. For the case of unit statistics, we need to estimate the population total $Y = \sum_{k\in U} y_{(u)k}$ by using auxiliary information, which can improve the estimation given by the simple two-stage design-weighted (Horvitz–Thompson) estimator

$$\widehat{Y}_\pi = \sum_{k\in s} d_k y_{(u)k} = \sum_{i\in s_I} d_{Ii} \left(\sum_{k\in s_i} d_{k|i} y_{(u)k} \right). \tag{1}$$

Note that, for variables defined for clusters, the total $Y = \sum_{k\in U} y_{(u)k}$ can be written as $Y_I = \sum_{i\in U_I} y_{(c)i}$.

We wish to estimate a population total. Many parameters can often be expressed as a function of two or more population totals, which can be estimated by replacing each unknown population total by its Horvitz–Thompson estimator or by the calibration estimator. Similarly, the problem of estimating alternative parameters related to the mean is straightforward when the

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

problem of estimating a total is known. For example, the estimation of a mean, $\overline{Y} = Y/N$, can be obtained by dividing the estimation of the total by the population size, $\widehat{\overline{Y}} = \widehat{Y}/N$, which is known in some surveys. A proportion P can be seen as a particular case of a mean of a dichotomous variable and then $\widehat{P} = \widehat{Y}/N$.

Auxiliary information can exist simultaneously at the unit and the cluster level. Let $\mathbf{x}_{(u)k}$ be an auxiliary vector value associated with the kth unit, and $\mathbf{x}_{(c)i}$ an auxiliary vector value associated with the ith cluster.

Note that single-stage sampling of clusters (households, companies, and schools) is followed by observation of all in-scope individuals in a selected cluster. Under these conditions, s is now the sample of individuals (i.e., the set of individuals in all selected clusters), and $d_k = d_{Ii}d_{k|i} = d_{Ii}$ (because of single stage cluster sampling).

Calibration (Deville and Särndal 1992) provides a systematic way to take auxiliary information into account. In many standard settings, the calibration provides a simple and practical approach to incorporating auxiliary information into the estimation stage. In our context, a calibration estimator for unit statistics based on the information $\mathbf{x}_{(u)k}$ is

$$\widehat{Y}_{\text{cal}} = \sum_{k \in s} w_k y_{(u)k}, \tag{2}$$

where the unit weights w_k satisfy

$$\sum_{k \in s} w_k \mathbf{x}_{(u)k} = \sum_{k \in U} \mathbf{x}_{(u)k} \tag{3}$$

and minimize $D(w_k, d_k)$ for a given distance D.

The auxiliary information can be obtained from different sources and it could be in a conflict, although it is not the common situation. For example, the auxiliary information at the unit level used in surveys of voting intentions is taken from the individual, by means of personal voting recall. If only this information is used for calibrating, the estimates for the auxiliary variable do not usually coincide with the results of the previous election. On the other hand, auxiliary information at the cluster level can be taken from the vote count. Thus, we calibrate using auxiliary information from both levels and avoid inconsistencies.

A calibration estimator for cluster statistics based on the information $\mathbf{x}_{(c)i}$ is

$$\widehat{Y}_{I,\text{cal}} = \sum_{i \in s_I} w_{Ii} y_{(c)i}, \tag{4}$$

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

where the cluster weights w_{Ii} satisfy

$$\sum_{i \in s_I} w_{Ii} \mathbf{x}_{(c)i} = \sum_{i \in U_I} \mathbf{x}_{(c)i} \tag{5}$$

and minimize $D(w_{Ii}, d_{Ii})$.

Estevao and Särndal (2006) also derive calibration estimators based on the combined information provided by the cluster totals and unit totals. In order to calibrate the combined auxiliary information, they proposed two options for integrated weighting: $w_{Ii} = \sum_{i \in s_i} w_k / N_i$ for every $i \in s_I$, where N_i is the known size of cluster i or $w_k = d_{k|i} w_{Ii}$ for the selected units k in cluster $i \in s_I$. The unit weight w_k is the same for all units within the cluster and equal to the calibrated weight w_{Ii} for the cluster.

Nonintegrated calibration is also possible, as follows: starting from d_{Ii}, compute cluster weights w_{Ii} for $i \in s_I$, calibrated to the cluster information. In an independent second calibration, starting from $d_k = d_{Ii} d_{k|i}$, compute unit weights w_k for $k \in s$ calibrated to the unit information. The resulting estimators are $\widehat{Y}_{I,\text{cal}} = \sum_{i \in s_I} w_{Ii} y_{(c)i}$ for cluster statistics and $\widehat{Y}_{\text{cal}} = \sum_{k \in s} w_k y_{(u)k}$ for unit statistics.

In the next section, we define an alternative estimator for unit statistics, using the same available auxiliary information as the estimators $\widehat{Y}_{I,\text{cal}}$ and $\widehat{Y}_{\text{cal}}$.

A Shrinkage Estimator for Unit Statistics

Let $\widehat{Y}_{\text{cal}} = \sum_{k \in s} w_k y_{(u)k}$ be the calibration estimator for the population total Y defined by equation (2), and let $\widehat{Y}_{I,\text{cal}} = \sum_{i \in s_I} w_{Ii} y_{(c)i}$ be the calibration estimator for the population total defined by equation (4), where the unit weights w_k and the cluster weights w_{Ii} satisfy, respectively, equations (3) and (5).

The calibration equations impose consistency on the weight system, so that, when applied to the auxiliary variables, it will confirm (be consistent with) known aggregates for those same auxiliary variables. The estimator $\widehat{Y}_{\text{cal}}$ incorporates unit-level auxiliary information, whereas the estimator $\widehat{Y}_{I,\text{cal}}$ incorporates cluster-level auxiliary information, but not in an integrated pattern. The calibration sets out to modify the initial weights d_k or d_{iI} into new weights w_k or w_{Ii}, determined to be close to the d_k or d_{Ii}. Then, we obtain nearly design unbiased estimates. Auxiliary information used through calibration improves the accuracy of survey estimates.

We propose an estimator based on composite information, as follows. Starting from d_{Ii}, compute cluster weights w_{Ii} for $i \in s_I$, calibrated to the

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

cluster information. After calibration, let $w_k^I = w_{Ii}$, for all k in the same i, for $i \in s_I$ and let $\widehat{Y}_{\text{cal}}^I = \sum_{k \in s} w_k^I y_{(u)k}$.

Shrinkage is a natural way to improve the available estimates, in terms of the MSE. For example, composite estimators are used in small area estimation to balance the potential bias of the synthetic estimator against the instability of the direct estimator (see Rao 2003). Shrinkage is explicit in James–Stein type inference (James and Stein 1961). The use of shrinkage estimators in the context of regression analysis has been described by Copas (1983). Copas (1993) also applies shrinkage in a context where the problem is to predict a binary response on the basis of binary explanatory variables. Similarly, Schäfer and Strimmer (2005) define a shrinkage estimator of the covariance matrix, and Rueda and Menéndez (2010) use shrinkage region in small area estimation.

The efficiency of using weighted estimators depends on the relationship between the auxiliary and the interest variables. An estimator based on the auxiliary information at the cluster level will be more efficient when this information is well related to the variable of interest. Similarly, one based on the auxiliary information at the unit level will be more efficient when this information is well related to the variable of interest. A natural way to balance the efficiency of these estimators is to take a weighted average.

Following Thompson (1968), we propose shrinking the unit estimator $\widehat{Y}_{\text{cal}}$ toward the estimator $\widehat{Y}_{\text{cal}}^I$. We thus obtain $\widetilde{Y} = K\widehat{Y}_{\text{cal}} + (1-K)\widehat{Y}_{\text{cal}}^I$, where K is a constant satisfying $0 < K < 1$.

An optimal choice of K can be calculated by minimizing the variance of , which is given by

$$V(\widetilde{Y}) = K^2 V(\widehat{Y}_{\text{cal}}) + (1-K)^2 V(\widehat{Y}_{\text{cal}}^I) + 2K(1-K)\text{Cov}(\widehat{Y}_{\text{cal}}, \widehat{Y}_{\text{cal}}^I).$$

As this equation is a quadratic equation of K, its sole extreme is found straightforwardly. The values of K that minimizes this variance are given by

$$K_{\text{opt}} = \frac{V(\widehat{Y}_{\text{cal}}^I) - \text{Cov}(\widehat{Y}_{\text{cal}}, \widehat{Y}_{\text{cal}}^I)}{V(\widehat{Y}_{\text{cal}}) + V(\widehat{Y}_{\text{cal}}^I) - 2\text{Cov}(\widehat{Y}_{\text{cal}}, \widehat{Y}_{\text{cal}}^I)}. \tag{6}$$

Note that the denominator in equation (6) is the variance of the difference $\widehat{Y}_{\text{cal}}^I - \widehat{Y}_{\text{cal}}$ and the numerator in equation (6) is the covariance of this difference with $\widehat{Y}_{\text{cal}}^I$.

K_{opt} can be used to define the optimum expression

$$\widetilde{Y}_{\text{opt}} = K_{\text{opt}}\widehat{Y}_{\text{cal}} + (1-K_{\text{opt}})\widehat{Y}_{\text{cal}}^{\text{I}}.$$

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

The variance of this estimator is given by:

$$V(\widetilde{Y}_{\text{opt}}) = V_{\min}(\widetilde{Y}) = \frac{V(\widehat{Y}^{\text{I}}_{\text{cal}})V(\widehat{Y}_{\text{cal}}) - \text{Cov}^2(\widehat{Y}_{\text{cal}}, \widehat{Y}^I_{\text{cal}})}{V(\widehat{Y}_{\text{cal}}) + V(\widehat{Y}^I_{\text{cal}}) - 2\text{Cov}(\widehat{Y}_{\text{cal}}, \widehat{Y}^I_{\text{cal}})}.$$

Sample-based Counterparts

The optimal coefficient K_{opt} depends on population variances and covariances, which are usually unknown in practice, and so $\widetilde{Y}_{\text{opt}}$ cannot be calculated. For this reason, these variances and covariances must be estimated.

The population-based quantities $V(\widehat{Y}^I_{\text{cal}})$, $V(\widehat{Y}_{\text{cal}})$ and $\text{Cov}(\widehat{Y}_{\text{cal}}, \widehat{Y}^I_{\text{cal}})$ have sample-based analogs, which are now defined using two different procedures.

In the case of $V(\widehat{Y}_{\text{cal}})$, automated linearization (see Estevao and Särndal 2006) identifies the linearized statistic and the residuals that determine the approximate variance,

$$V(\widehat{Y}_{\text{cal}}) \approx V\left(\sum_{k\in s} d_k e_k\right) = \sum_{k,l\in U} F_{kl} e_k e_l,$$

with $F_{kl} = (d_k d_l / d_{kl}) - 1$ for $k \neq l$ and $F_{kk} = d_k - 1$ for $l = k$ and where $e_k = y_{(u)k} - \mathbf{x}^t_{(u)k}\mathbf{B}$ and $\mathbf{B} = \left(\sum_{k\in U} \mathbf{x}_{(u)k}\mathbf{x}^t_{(u)k}\right)^{-1} \sum_{k\in U} \mathbf{x}_{(u)k} y_{(u)k}$. Finally, the variance can be estimated by

$$\hat{V}(\hat{Y}_{\text{cal}}) = \sum_{k\in s}\sum_{\ell\in s}(d_k d_l - d_{kl})\widehat{e}_k\widehat{e}_\ell,$$

where $\widehat{e}_k = y_{(u)k} - \mathbf{x}^t_{(u)k}\widehat{\mathbf{B}}$ and $\widehat{\mathbf{B}} = \left(\sum_{k\in s} d_k \mathbf{x}_{(u)k}\mathbf{x}^t_{(u)k}\right)^{-1} \sum_{k\in s} d_k \mathbf{x}_{(u)k} y_{(u)k}$.

In the case of $V(\widehat{Y}^I_{\text{cal}})$, since $\widehat{Y}^I_{\text{cal}}$ is an estimator of the population total, automated linearization yields

$$\hat{V}(\hat{Y}^I_{\text{cal}}) = \sum_{i\in s_I}\sum_{j\in s_I}(d_{Ii} d_{Ij} - d_{Iij})\widehat{e}_{Ii}\widehat{e}_{Ij},$$

where $\widehat{e}_{Ii} = y_{(c)i} - \mathbf{x}^t_{(c)i}\widehat{\mathbf{B}}_I$ and $\widehat{\mathbf{B}}_I = \left(\sum_{i\in s_I} d_{Ii}\mathbf{x}_{(c)i}\mathbf{x}^t_{(c)i}\right)^{-1} \sum_{i\in s_I} d_{Ii}\mathbf{x}_{(c)i} y_{(c)i}$.

In the case of $\text{Cov}(\widehat{Y}_{\text{cal}}, \widehat{Y}^I_{\text{cal}})$, we have

$$\text{Cov}(\widehat{Y}_{\text{cal}}, \widehat{Y}^I_{\text{cal}}) \approx \text{Cov}\left(\sum_{k\in s} d_k e_k, \sum_{i\in s_I} d_{Ii} e_{Ii}\right) = \text{Cov}\left(\sum_{i\in s_I} d_{Ii} e^I_i, \sum_{i\in s_I} d_{Ii} e_{Ii}\right).$$

Note that $d_k = d_{Ii} d_{k|i} = d_{Ii}$ due to the single-stage cluster sampling. The residuals e^I_i are determined by first calculating e_k based on a regression of

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

$y_{(u)k}$ on $\mathbf{x}^t_{(u)k}$ at the unit level and then summing the e_k within each cluster to produce $e^I_i = \sum_{k \in U_i} e_k$.

Finally, the covariance can be estimated by

$$\widehat{\text{Cov}}\left(\widehat{Y}_{\text{cal}}, \widehat{Y}^I_{\text{cal}}\right) = \sum_{i \in s_I} \sum_{j \in s_I} (d_{Ii} d_{Ij} - d_{Iij}) \widehat{e}^I_i \widehat{e}_{Ij} \tag{7}$$

The following optimum estimator can be defined

$$\widehat{Y}_{\text{opt}_D} = \widehat{K}_{\text{opt}_D} \widehat{Y}_{\text{cal}} + (1 - \widehat{K}_{\text{opt}_D}) \widehat{Y}^I_{\text{cal}},$$

where $\widehat{K}_{\text{opt}_D}$ denotes that estimates are substituted for the variances and covariances in equation (6).

The calibration approach can be readily adapted to the estimation of more complex parameters than a population total. For example, it can be used to estimate the population variance (Särndal 1982; Singh 2001), quantiles and distribution functions (Kovaĉević 1997; Rueda et al. 2007; Tillé 2002), and other complex parameters (Krapavickaite and Plikusas 2005; Plikusas 2006). It is also possible to obtain calibration estimators for variances and covariances of calibration estimators. In our case, the unknown population quantities $V(\widehat{Y}^I_{\text{cal}})$, $V(\widehat{Y}_{\text{cal}})$, and $\text{Cov}(\widehat{Y}_{\text{cal}}, \widehat{Y}^I_{\text{cal}})$ can be estimated as follows (see Singh 2001, 2004, 2010; Singh et al. 1999):

$$\begin{aligned}
\hat{V}_W(\hat{Y}_{\text{cal}}) &= \sum_{k \in s} \sum_{\ell \in s} (\pi_{k\ell} - \pi_k \pi_\ell)/\pi_{k\ell} w_k \widehat{e}_k w_\ell \widehat{e}_\ell \\
\hat{V}_W(\hat{Y}^I_{\text{cal}}) &= \sum_{i \in s_I} \sum_{j \in s_I} (\pi_{Iij} - \pi_{Ii} \pi_{Ij})/\pi_{Iij} w_{Ii} \widehat{e}_{Ii} w_{Ij} \widehat{e}_{Ij} \\
\widehat{\text{Cov}}_W(\widehat{Y}_{\text{cal}}, \widehat{Y}^I_{\text{cal}}) &= \sum_{k \in s_I} \sum_{\ell \in s_I} (\pi_{Iij} - \pi_{Ii} \pi_{Ij})/\pi_{Iij} w_{Ii} \widehat{e}^I_i w_{Ij} \widehat{e}_{Ij}
\end{aligned} \tag{8}$$

A second optimum estimator can be defined as

$$\widehat{Y}_{\text{opt}_W} = \widehat{K}_{\text{opt}_W} \widehat{Y}_{\text{cal}} + (1 - \widehat{K}_{\text{opt}_W}) \widehat{Y}^I_{\text{cal}},$$

where $\widehat{K}_{\text{opt}_W}$ denotes that these calibration estimates are substituted for the variances and covariances in equation (6).

We now consider the problem of estimating the variances of the two proposed estimators $\widehat{Y}_{\text{opt}_D}$ and $\widehat{Y}_{\text{opt}_W}$. This is not a simple issue because the factors $\widehat{K}_{\text{opt}_D}$ and $\widehat{K}_{\text{opt}_W}$ are derived from survey data.

One way to address this problem is to consider resampling methods such as random groups, balanced half samples, jackknife, or bootstrap techniques

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

(see Wolter 2007 or Gershunskaya, Jiang, and Lahiri 2009 for a detailed review of these techniques).

Another option is to use asymptotic results. Randles (1982) studied the effects on convergence of substituting parameter estimates into a statistics. This important result (which has been used by authors such as Chambers and Dustan 1986;Gijbels and Veraverbeke 1988; Rao et al. 1990; and Rueda et al. 2007) assumes that the sample observations can be treated as independent and identically distributed realizations. Recently, Wang and Opsomer (2011) extended this result to a general sampling design.

In this article, we follow the framework given by Isaki and Fuller (1982) in which the properties of estimators are established under a fixed sequence of populations and a corresponding sequence of random sampling designs. Therefore, we assume an increasing sequence of finite populations $\{U_N\}$ with $N \to \infty$.

Using notation given by Randles (1982), the proposed estimator is denoted by $\widehat{Y}_{\text{opt}_D} = T_n(\widehat{K}_{\text{opt}_D}) = T_n(\widehat{\lambda})$ and $\widetilde{Y}_{\text{opt}} = T_n(K_{\text{opt}}) = T_n(\lambda)$, where the estimator $\widehat{\lambda}$ consistently estimates λ. If the sampling design verifies the assumptions given by Wang and Opsomer (2011, Section 2), then theorem 1 given by Wang and Opsomer (2011) can be applied, because the estimator $\widehat{\lambda}$ and the function T verify all the necessary assumptions. We thus conclude that the asymptotic distribution of $T_n(\widehat{\lambda})(=\widehat{Y}_{\text{opt}_D})$ is the same as that of $T_n(\lambda)(=\widetilde{Y}_{\text{opt}})$.

Thus, the asymptotic variance of the estimator $\widehat{Y}_{\text{opt}_D}$ is given by:

$$\text{AV}(\widehat{Y}_{\text{opt}_D}) = V_{\min}(\widetilde{Y}_{\text{opt}}) = \frac{V(\widehat{Y}^I_{\text{cal}})V(\widehat{Y}_{\text{cal}}) - \text{Cov}^2(\widehat{Y}_{\text{cal}}, \widehat{Y}^I_{\text{cal}})}{V(\widehat{Y}_{\text{cal}}) + V(\widehat{Y}^I_{\text{cal}}) - 2\text{Cov}(\widehat{Y}_{\text{cal}}, \widehat{Y}^I_{\text{cal}})}.$$

An estimator for this approximative variance is as follows:

$$\widehat{V}(\widehat{Y}_{\text{opt}_D}) = \frac{\widehat{V}(\widehat{Y}^{\text{I}}_{\text{cal}})\widehat{V}(\widehat{Y}_{\text{cal}}) - \widehat{\text{Cov}}^2(\widehat{Y}_{\text{cal}}, \widehat{Y}^I_{\text{cal}})}{\widehat{V}(\widehat{Y}_{\text{cal}}) + \widehat{V}(\widehat{Y}^I_{\text{cal}}) - 2\widehat{\text{Cov}}(\widehat{Y}_{\text{cal}}, \widehat{Y}^I_{\text{cal}})}.$$

The variance estimator of $\widehat{Y}_{\text{opt}_W}$ can be similarly defined.

Simulation Study

In this section, the empirical performance of the proposed estimators is evaluated via a simulation study.

The aim of this is to use auxiliary information to find better alternatives for estimating the population total $Y = \sum_{k \in U} y_{(u)k}$ than the single-stage design-weighted estimator $\widehat{Y}_{\pi}$ defined by equation (1). For this purpose,

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

we use auxiliary information, which is available for both units and clusters. Another approach is that suggested by Lemaître and Dufour (1987), which uses the same auxiliary information: A set of calibrated weights is computed by modifying the design weights through the available auxiliary information. Alternatively, Estevao and Särndal's (2006) approach obtains exactly the same individual weights. We compare the proposed estimators with the one proposed by Estevao and Särndal (2006), denoted by $\widehat{Y}_{\text{ES}}$.

Observe that K_{opt} depends on the unknown quantities $V(\widehat{Y}_{\text{cal}})$, $V(\widehat{Y}^{I}_{\text{cal}})$, and $\text{Cov}(\widehat{Y}_{\text{cal}}, \widehat{Y}^{I}_{\text{cal}})$. Three estimators are considered for K_{opt}. First, the variances and the covariance are estimated in overall simulation runs, and the resulting estimator is noted by $\widehat{Y}_{\text{opt}_S}$. K_{opt_D} and K_{opt_W} are also considered, which produce the estimators $\widehat{Y}_{\text{opt}_D}$ and $\widehat{Y}_{\text{opt}_W}$. Note that $\widehat{Y}_{\text{opt}_S}$ is computed merely for the sake of comparison, since this estimator cannot be obtained in practice.

The first population used in this simulation study was that of the PISA. This program was developed to evaluate students aged 15 years, involved in schoolwide educational and technical programs from countries belonging to the OECD system, as well as other associated partners. The data analyzed corresponded to the year 2006, for 57 countries, and focused on reading skills, mathematics, and science.

The OECD website provides microdata on the 2006 study, including information on students, families, and schools, while ensuring the participants' anonymity. The simulation study in this article is based on the OECD data for Spain.

The PISA-SPAIN microdata report contains information on the tests conducted at 686 schools with the participation of 19,604 students. We considered the units with no missing data, thus obtaining a population with size $N = 18{,}341$ students (units) grouped into $N_I = 673$ schools (cluster).

For the PISA population, we chose two variables of interest, a qualitative dichotomous variable (Sci. future—After secondary school) and a quantitative variable (plausible value in math). We chose gender (male and female), educational level of mother, educational level of father, and the highest educational level of parents (no studies, primary, secondary, and higher education) as the auxiliary information for units. Finally, we considered the type of school (public or private) and the kind of school community (village or small town, town, city, or large city) as the auxiliary information for clusters.

The second finite population used in the simulation was obtained from the Household Budget Survey (HBS). This is a survey of a representative random sample of all private households in Spain. Such surveys have been carried out periodically since 2006. The web page of the Spanish National Statistics Institute (INE) includes statistics that can be used to obtain

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

microdata files, each of which contains individual data for a given statistic, filtered appropriately to make the information anonymous and thus ensure confidentiality.

The HBS-2006 microdata contain 55,699 units grouped into 19,435 clusters. We considered only the units with no missing data, and thus obtained a population with size $N = 9{,}243$ individuals grouped into $N_I = 5{,}800$ households. The main features of the HBS-2006 can be consulted on the INE web page.

In this case, the variable of interest is the income of the HBS population. We considered gender (male and female) and educational level (no studies, primary, secondary, and higher education) as the auxiliary information for the units. Finally, we considered the educational level of the main breadwinner (primary, secondary, and university) as the auxiliary information for the clusters (households).

Two simulation studies were performed. In the first case, for each simulation, a simple random sample of size m of primary stage units (schools and households, respectively) was drawn from the cluster population. Then, all units in the selected cluster were observed and a sample of size n of secondary stage units was obtained (individuals and students, respectively). In the second case, the Midzuno's (1952) method was used to select a sample of units (schools or households) with unequal probabilities (proportional to cluster size).

There are powerful arguments for using calibration estimators, such as their excellent asymptotic properties. In the distance minimization method, the distance function can be formulated so that negative weights are excluded, while still satisfying the given calibration equations. Some programs minimize the distance, subject to calibration constraints and individual bounds on the weights. Many practitioners support the reasonable requirement that all weights be positive (even greater than unity) and that very large weights should be avoided. Computational aspects, extreme weights, and outliers have been discussed in various articles. The user's guide should be consulted for each program to see exactly how the computational issues are handled.

The proposed estimators use weights that assume restrictions separately, which implies that extreme weights are unlikely to appear. A detailed discussion of this computational aspect regarding the calibration method can be found in Silva (2003). In any case, such situations arise from the calibration method, and shrinkage is not responsible for causing extreme weights.

The weights of calibration estimators can be computed by means of different calibration methods, depending on the measure distance selected. For comparative purposes, it is interesting to show how accuracy was achieved when either of the calibration methods was used. In our simulation study,

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

three calibration methods were included: linear, raking, and logit (see Deville and Särndal 1992).

The Horvitz–Thompson estimator $\widehat{Y}_{\pi}$, the Estevao–Särndal estimator $\widehat{Y}_{\mathrm{ES}}$ and the proposed estimators, $\widehat{Y}_{\mathrm{opt}_S}$, $\widehat{Y}_{\mathrm{opt}_D}$, $\widehat{Y}_{\mathrm{opt}_W}$ were computed from each sample. In the case of a dichotomous variable, the population proportion P was estimated by $\widehat{P} = \widehat{Y}/N$ for each estimator of the population total $\widehat{Y}$. This process was repeated $B = 1{,}000$ times.

The above simulation was then repeated with $m = 25$, 50, 75, 150, 200, and 250 for the HBS population. Sample sizes from $n = 25$ to $n = 250$ were used, and thus sampling fractions from $f = 25/5{,}800 = 0.004$ to $f = 250/5{,}800 = 0.04$ were considered. Similarly, we considered $m = 20$, 25, 30, 40, 45, and 50 for the PISA population, and the sampling fractions ranged from $f = 20/673 = 0.0297$ to $f = 50/673 = 0.0743$ in this case.

The performance of each total estimator was measured and compared in terms of relative bias (RB) and relative efficiency (RE). The simulated values of RB and RE for a particular total estimator T were computed as

$$RB = B^{-1}\sum_{b=1}^{B}(T^b - Y)/Y, \qquad RE = \mathrm{MSE}(\widehat{Y}_{\pi})/\mathrm{MSE}(T^b)$$

where $\mathrm{MSE}(T^b) = B^{-1}\sum_{b=1}^{B}(T^b - Y)^2$, $\mathrm{MSE}(\widehat{Y}_{\pi}) = B^{-1}\sum_{b=1}^{B}(\widehat{Y}_{\pi}^b - Y)^2$, and T^b and Y_{π}^b are the values of T and $\widehat{Y}_{\pi}$ from the bth simulation, respectively.

Tables 1 to 3 show the results obtained under simple random sampling. In Table 1, the main variable is a dichotomic variable, Sci. future, and the school-auxiliary variables are type of school and kind of community; the student-auxiliary variables are gender, mother's educational level, father's educational level, and student's highest educational level. Note that a proportion is estimated in this case.

In Table 2, the main variable is Plausible value in mathematics (a continuous variable); and the auxiliary information is the same as in Table 1. In Table 3, the main variable is income in HBS population; the household-auxiliary variable is main breadwinner's educational level; and the person-auxiliary variables are gender and educational level.

Tables 1 to 3 show that the three calibration methods (linear, raking, and logit) give similar results in terms of RB and RE. Therefore, the three methods lead to the same conclusions. It can be seen that the Estevao–Särndal estimator is always more efficient than the Horvitz–Thompson estimator, and moreover that the three optimal estimators give larger values RE than does the Estevao–Särndal estimator.

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

Table 1. Values of RB% and RE% for the Various Estimators and the PISA-SPAIN Population.

			Linear		Ranking		Logit	
	m	$\bar{n}$	RB%	RE%	RB%	RE%	RB%	RE%
$\widehat{P}_{\pi}$	20	544.31	−.227	100.00	−.227	100.00	−.227	100.00
$\widehat{P}_{ES}$			−.409	115.62	−.117	104.42	−.051	−103.47
$\widehat{P}_{opt_S}$			−.190	162.13	−.194	162.13	−.194	162.13
$\widehat{P}_{opt_W}$			−.250	148.24	−.245	148.10	−.246	148.15
$\widehat{P}_{opt_D}$			−.154	148.02	−.149	147.97	−.150	147.97
$\widehat{P}_{\pi}$	25	680.80	.242	100.00	.242	100.00	.242	100.00
$\widehat{P}_{ES}$			.027	127.58	.081	114.61	.125	115.10
$\widehat{P}_{opt_S}$			.245	179.28	.238	180.12	.237	180.12
$\widehat{P}_{opt_W}$			.250	151.13	.251	150.90	.251	150.94
$\widehat{P}_{opt_D}$			.317	151.79	.316	151.68	.316	151.68
$\widehat{P}_{\pi}$	30	817.03	.079	100.00	.079	100.00	.079	100.00
$\widehat{P}_{ES}$			−.108	143.35	−.105	125.96	−.059	127.81
$\widehat{P}_{opt_S}$			.077	178.73	.078	179.02	.078	179.02
$\widehat{P}_{opt_W}$			.072	160.41	.073	160.21	.073	160.23
$\widehat{P}_{opt_D}$			.132	160.00	.132	159.95	.132	159.95
$\widehat{P}_{\pi}$	35	953.69	−.388	100.00	−.388	100.00	−.388	100.00
$\widehat{P}_{ES}$			−.533	144.55	−.498	135.23	−.502	135.85
$\widehat{P}_{opt_S}$			−.424	168.52	−.423	168.55	−.423	168.55
$\widehat{P}_{opt_W}$			−.407	153.07	−.404	153.02	−.405	153.05
$\widehat{P}_{opt_D}$			−.361	153.54	−.361	153.52	−.361	153.52
$\widehat{P}_{\pi}$	40	1,089.15	−.063	100.00	−.063	100.00	−.063	100.00
$\widehat{P}_{ES}$			.006	152.95	.101	145.99	.122	144.18
$\widehat{P}_{opt_S}$			.015	179.28	.014	178.92	.013	178.89
$\widehat{P}_{opt_W}$			.009	160.90	.009	160.95	.009	160.95
$\widehat{P}_{opt_D}$			.058	160.05	.058	160.05	.058	160.05
$\widehat{P}_{\pi}$	45	1,227.41	−.035	100.00	−.035	100.00	−.035	100.00
$\widehat{P}_{ES}$			.034	153.70	.021	144.45	.089	146.84
$\widehat{P}_{opt_S}$			−.076	174.40	−.075	174.28	−.075	174.28
$\widehat{P}_{opt_W}$			−.044	160.31	−.003	160.26	−.044	160.28
$\widehat{P}_{opt_D}$			−.003	160.62	−.361	160.62	−.003	160.62
$\widehat{P}_{\pi}$	50	1,363.28	.072	100.00	.072	100.00	.072	100.00
$\widehat{P}_{ES}$			−.172	150.44	−.140	143.80	−.134	148.72
$\widehat{P}_{opt_S}$			−.068	166.25	−.070	166.56	−.070	166.56
$\widehat{P}_{opt_W}$			−.033	154.42	−.033	154.32	−.033	154.34
$\widehat{P}_{opt_D}$			.003	154.15	.003	154.15	.003	154.15

Note: m Denotes the size of primary stage units and $\bar{n}$ the average size over $B = 1{,}000$ simulations of final stage units. Linear, raking, and logit are used as calibration methods. Main variable: Sci. future. School-auxiliary variables: Type of school and kind of community. Student-auxiliary variables: gender, mother educational level, father educational level, and highest educational level. Simple random sampling is considered.

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

Table 2. Values of RB% and RE% for the Various Estimators and the PISA-SPAIN Population.

	m	$\bar{n}$	Linear		Ranking		Logit	
			RB%	RE%	RB%	RE%	RB%	RE%
$\widehat{Y}_{\pi}$	20	543.57	−.289	100.00	−.289	100.00	−.289	100.00
$\widehat{Y}_{ES}$			−.028	87.95	−.037	78.37	−.018	77.45
$\widehat{Y}_{opt_S}$			−.012	127.55	−.017	128.20	−.017	128.20
$\widehat{Y}_{opt_W}$			.000	125.15	.001	125.15	.001	125.15
$\widehat{Y}_{opt_D}$			.004	125.94	.005	125.78	.005	125.78
$\widehat{Y}_{\pi}$	25	681.67	.117	100.00	.117	100.0	.117	100.00
$\widehat{Y}_{ES}$			.054	88.18	.039	82.71	.028	84.60
$\widehat{Y}_{opt_S}$			.047	104.60	.062	127.06	.062	127.06
$\widehat{Y}_{opt_W}$			.092	117.78	.092	117.64	.092	117.64
$\widehat{Y}_{opt_D}$			.096	117.50	.096	117.50	.096	117.50
$\widehat{Y}_{\pi}$	30	818.74	.042	100.00	.042	100.00	.042	100.00
$\widehat{Y}_{ES}$			−.108	95.51	−.095	91.32	−.094	92.50
$\widehat{Y}_{opt_S}$			−.101	112.23	−.103	116.95	−.103	116.95
$\widehat{Y}_{opt_W}$			−.076	115.74	−.076	115.74	−.076	115.74
$\widehat{Y}_{opt_D}$			−.073	115.74	−.073	115.74	−.073	115.74
$\widehat{Y}_{\pi}$	35	953.81	.022	100.00	.022	100.00	.022	100.00
$\widehat{Y}_{ES}$			.018	107.41	.019	103.19	.011	105.59
$\widehat{Y}_{opt_S}$			.021	122.39	.024	123.30	.024	123.30
$\widehat{Y}_{opt_W}$			.038	121.21	.038	121.21	.038	121.21
$\widehat{Y}_{opt_D}$			.042	121.21	.042	121.21	.042	121.21
$\widehat{Y}_{\pi}$	40	1,091.05	.086	100.00	.086	100.00	.086	100.00
$\widehat{Y}_{ES}$			−.016	114.81	−.003	113.63	−.014	115.20
$\widehat{Y}_{opt_S}$			−.002	129.19	−.003	133.51	003	133.51
$\widehat{Y}_{opt_W}$			.015	129.19	.015	129.19	.015	129.19
$\widehat{Y}_{opt_D}$			.019	128.70	.019	128.70	.019	128.70
$\widehat{Y}_{\pi}$	45	1,227.10	.053	100.00	.053	100.00	.053	100.00
$\widehat{Y}_{ES}$			.018	105.70	.006	100.20	.009	105.04
$\widehat{Y}_{opt_S}$			−005	121.35	−004	123.00	−004	123.00
$\widehat{Y}_{opt_W}$			.006	122.10	.006	122.24	.006	122.24
$\widehat{Y}_{opt_D}$			.010	121.95	.010	121.95	.010	121.95
$\widehat{Y}_{\pi}$	50	1,362.21	−.031	100.00	−.031	100.00	−.031	100.00
$\widehat{Y}_{ES}$			−.029	116.27	−.024	110.13	−.036	118.34
$\widehat{Y}_{opt_S}$			−.005	128.04	−.004	128.53	−.004	128.53
$\widehat{Y}_{opt_W}$			.000	126.10	.000	126.10	.000	126.10
$\widehat{Y}_{opt_D}$			.003	126.10	.003	126.10	.003	126.10

Note: m Denotes the size of primary stage units and $\bar{n}$ the average size over $B = 1{,}000$ simulations of final stage units. Linear, raking, and logit are used as calibration methods. Main variable: Plausible value in mathematics. School-auxiliary variables: Type of school and type of community. Student-auxiliary variables: gender, mother educational level, father educational level, and highest educational level. Simple random sampling is considered.

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

Table 3. Values of RB% and RE% for the Various Estimators and the HBS Population.

			Linear		Ranking		Logit	
	m	$\bar{n}$	RB%	RE%	RB%	RE%	RB%	RE%
$\widehat{Y}_{\pi}$	25	40.14	1.92	100.00	1.92	100.00	1.92	100.00
$\widehat{Y}_{ES}$			.605	155.86	.500	152.00	.503	152.53
$\widehat{Y}_{opt_S}$			.471	197.47	.460	198.49	.461	198.41
$\widehat{Y}_{opt_W}$			.106	193.54	.106	193.61	.105	193.61
$\widehat{Y}_{opt_D}$			.676	189.93	.676	189.93	.676	189.93
$\widehat{Y}_{\pi}$	50	79.93	.070	100.00	.070	100.00	.070	100.00
$\widehat{Y}_{ES}$			.223	203.21	.129	204.46	.135	204.62
$\widehat{Y}_{opt_S}$			−.013	231.43	−.010	231.27	−.011	231.32
$\widehat{Y}_{opt_W}$			−.394	226.35	−.395	226.30	−.395	226.30
$\widehat{Y}_{opt_D}$			.002	226.30	.002	226.30	.002	226.30
$\widehat{Y}_{\pi}$	75	119.46	−.086	100.00	−.086	100.00	−.086	100.00
$\widehat{Y}_{ES}$			.146	208.94	.034	210.13	.049	210.17
$\widehat{Y}_{opt_S}$			.007	225.73	.006	225.53	.006	225.58
$\widehat{Y}_{opt_W}$			−.309	223.26	−.309	223.21	−.309	223.21
$\widehat{Y}_{opt_D}$			−.007	224.92	−.007	224.92	−.007	224.92
$\widehat{Y}_{\pi}$	100	159.54	.418	100.00	.418	100.00	.418	100.00
$\widehat{Y}_{ES}$			.276	209.95	.227	205.80	.238	206.19
$\widehat{Y}_{opt_S}$			.174	223.16	.174	223.02	.174	223.02
$\widehat{Y}_{opt_W}$			−.029	225.12	−.029	225.12	−.029	225.12
$\widehat{Y}_{opt_D}$			.174	222.47	.174	222.47	.174	222.47
$\widehat{Y}_{\pi}$	150	239.36	.084	100.00	.084	100.00	.084	100.00
$\widehat{Y}_{ES}$			.101	211.82	.051	212.86	.064	212.77
$\widehat{Y}_{opt_S}$			.039	216.97	.031	217.06	.031	217.06
$\widehat{Y}_{opt_W}$			−.108	217.96	−.108	217.96	−.108	217.96
$\widehat{Y}_{opt_D}$			.032	215.38	.032	215.38	.032	215.38
$\widehat{Y}_{\pi}$	200	318.87	.117	100.00	.117	100.00	.117	100.00
$\widehat{Y}_{ES}$			.097	217.72	.060	218.67	.061	220.60
$\widehat{Y}_{opt_S}$			.041	225.63	.042	225.53	.042	225.53
$\widehat{Y}_{opt_W}$			−.055	222.67	−.055	222.67	−.055	222.67
$\widehat{Y}_{opt_D}$			.040	223.11	.040	223.11	.040	223.11
$\widehat{Y}_{\pi}$	250	398.39	−.102	100.00	−.102	100.00	−.102	100.00
$\widehat{Y}_{ES}$			−.012	220.51	−.048	220.65	−.042	220.26
$\widehat{Y}_{opt_S}$			−.054	222.57	−.056	222.62	−.056	222.62
$\widehat{Y}_{opt_W}$			−.127	220.70	−.127	220.75	−.127	220.75
$\widehat{Y}_{opt_D}$			−.058	221.39	−.058	221.39	−.058	221.39

Note: m Denotes the size of primary stage units and $\bar{n}$ the average size over $B = 1{,}000$ simulations of final stage units. Linear, raking, and logit are used as calibration methods. Main variable: Income. Household-auxiliary variable: Main breadwinner educational level. Person-auxiliary variables: gender and educational level. Simple random sampling is considered.

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

There is no clear relationship between $\widehat{Y}_{\text{opt}_D}$ and $\widehat{Y}_{\text{opt}_W}$ in terms of RB and RE, although the estimator $\widehat{Y}_{\text{opt}_W}$ seems to be slightly more efficient than $\widehat{Y}_{\text{opt}_D}$.

For moderate sample sizes, the proposed estimator is clearly the most efficient estimator, and the gain in efficiency is larger for small sample sizes. The proposed and the Estevao–Särndal estimators give similar values of RE as the sample size increases.

Note that $\widehat{Y}_{\text{opt}}$ has the added disadvantage of having to estimate K_{opt}, although the effort required is not so great for moderate sample sizes and may be worth making, precisely in these cases, thus achieving a more efficient estimator.

Tables 4 to 6 show the results obtained under probability proportional to size sampling. The variables in Table 4 are the same as in Table 1, those in Table 5 are the same as in Table 2, and those in Table 6 are the same as in Table 3 plus age (as a continuous variable) as a person-auxiliary variable.

Tables 4 to 6 only include the design optimum estimator $\widehat{Y}_{\text{opt}_D}$. $\widehat{Y}_{\text{opt}_W}$ gives similar results to $\widehat{Y}_{\text{opt}_D}$. For reasons of space, $\widehat{Y}_{\text{opt}_W}$ and $\widehat{Y}_{\text{opt}_S}$ are omitted.

In Tables 4 and 5 (PISA population), we observe that the Estevao–Särndal estimator is not always more efficient than the Horvitz–Thompson estimator (only for $m = 50$ and the linear method is the Estevao–Särndal estimator more efficient than Horvitz–Thompson), whereas Table 6 (HBS population) shows that the Estevao–Särndal estimator is more efficient than Horvitz–Thompson except for $m = 20$ and with the raking and logit methods.

The proposed estimator is clearly the most efficient estimator, and the gain in efficiency with respect to the Estevao–Särndal estimators is larger for small sample sizes (Tables 4–6). This efficiency gain is moderate (RE $\simeq 102$) when estimating a population total (a quantitative variable, Table 5) and increases (RE varies from 116 to 118) when estimating a population proportion (a qualitative variable, Table 4), whereas RE ranges from 132 to 147 when the total income in the HBS population is estimated (Table 6).

From Tables 4 to 6, as in the case of simple random sampling, under the Midzuno sampling scheme (unequal probabilities), we observe that the optimal proposed estimator yields larger RE values than those of the Estevao–Särndal and Horvitz–Thompson estimators. The three calibration methods (linear, raking, and logit) produce the same conclusions. Finally, the gain in efficiency with respect to the Estevao–Särndal estimator increases as the sample size decreases.

Moreover, assuming unequal probabilities, the Estevao–Särndal estimator is not always more accurate than Horvitz–Thompson.

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

Table 4. Values of RB% and RE% for the Various Estimators and the PISA-SPAIN Population.

	m	$\bar{n}$	Linear		Ranking		Logit	
			RB%	RE%	RB%	RE%	RB%	RE%
$\widehat{P}_{\pi}$	20	564.438	−.173	100	−.173	100	−.173	100
$\widehat{P}_{ES}$			−.109	75.89	.113	69.99	.136	68.21
$\widehat{P}_{opt_D}$			−.172	116.05	−.173	116.07	−.172	116.08
$\widehat{P}_{\pi}$	25	707.461	.142	100	.142	100	.142	100
$\widehat{P}_{ES}$			−.111	80.03	−.011	70.55	.041	70.32
$\widehat{P}_{opt_D}$			.054	114.25	.054	114.27	.054	114.27
$\widehat{P}_{\pi}$	30	848.861	−.487	100	−.487	100	−.487	100
$\widehat{P}_{ES}$			−.352	89.23	−.304	83.99	−.298	83.86
$\widehat{P}_{opt_D}$			−.350	117.29	−.350	117.38	−.350	117.37
$\widehat{P}_{\pi}$	35	991.364	−.318	100	−.318	100	−.318	100
$\widehat{P}_{ES}$			−.377	86.01	−.361	79.11	−.425	79.65
$\widehat{P}_{opt_D}$			−.359	115.09	−.358	115.09	−.359	115.09
$\widehat{P}_{\pi}$	40	1,135.8	.030	100	.030	100	.030	100
$\widehat{P}_{ES}$			.180	95.44	.238	88.43	.231	89.55
$\widehat{P}_{opt_D}$			.049	116.34	.049	116.35	.049	116.35
$\widehat{P}_{\pi}$	45	1,275.921	−.005	100	−.005	100	−.005	100
$\widehat{P}_{ES}$			−.075	92.51	−.016	87.70	−.005	88.56
$\widehat{P}_{opt_D}$			−.052	114.93	−.052	114.92	−.052	114.92
$\widehat{P}_{\pi}$	50	1,416.123	−.321	100	−.321	100	−.321	100
$\widehat{P}_{ES}$			−.150	100.94	−.133	94.59	−.092	99.12
$\widehat{P}_{opt_D}$			−.217	118.46	−.218	118.45	−.218	118.45

Note: m Denotes the size of primary stage units and $\bar{n}$ the average size over $B = 1{,}000$ simulations of final stage units. Linear, raking, and logit are used as calibration methods. Main variable: Sci. future. School-auxiliary variables: Type of school and kind of community. Student-auxiliary variables: gender, mother educational level, father educational level, and highest educational level. Probability proportional to size sampling is considered.

Conclusion

This article discusses a new method to incorporate sampling weights for the data obtained from complex sampling designs, such as educational surveys with students grouped by schools. The approach suggested uses auxiliary information through calibration estimators. Two estimators are defined by shrinking a student-calibration estimator toward a school-calibration estimator.

The estimators proposed are suitable for data users who estimate totals or proportions with small school sample sizes, and efficient results can be obtained. The typical data user has access to the free software described here,

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

Table 5. Values of RB% and RE% for the Various Estimators and the PISA-SPAIN Population.

	m	$\bar{n}$	Linear		Ranking		Logit	
			RB%	RE%	RB%	RE%	RB%	RE%
$\widehat{Y}_{\pi}$	20	563.957	.021	100	.021	100	.021	100
$\widehat{Y}_{ES}$			.051	69.76	.028	67.14	.018	65.79
$\widehat{Y}_{opt_D}$			.014	102.29	.014	102.28	.014	102.28
$\widehat{Y}_{\pi}$	25	706.662	.002	100	.002	100	.002	100
$\widehat{Y}_{ES}$			−.065	76.43	−.079	70.31	−.085	69.20
$\widehat{Y}_{opt_D}$			.001	102.17	.001	102.15	.001	102.15
$\widehat{Y}_{\pi}$	30	848.241	−.101	100	−.101	100	−.101	100
$\widehat{Y}_{ES}$			−.114	76.70	−.103	70.25	−.116	72.25
$\widehat{Y}_{opt_D}$			−.101	101.61	−.101	101.63	−.101	101.62
$\widehat{Y}_{\pi}$	35	991.274	−.019	100	−.019	100	−.019	100
$\widehat{Y}_{ES}$			−.046	86.77	−.058	76.45	−.069	78.05
$\widehat{Y}_{opt_D}$			−.022	102.08	−.022	102.07	−.022	102.07
$\widehat{Y}_{\pi}$	40	1,136.184	.004	100	.004	100	.004	100
$\widehat{Y}_{ES}$			.033	86.07	.037	78.72	.028	82.00
$\widehat{Y}_{opt_D}$			.004	102.68	.004	102.68	.004	102.68
$\widehat{Y}_{\pi}$	45	1,275.299	.056	100	.056	100	.056	100
$\widehat{Y}_{ES}$			.034	84.84	.045	80.34	.040	82.14
$\widehat{Y}_{opt_D}$			.053	101.51	.053	101.50	.053	101.50
$\widehat{Y}_{\pi}$	50	1,415.947	−.016	100	−.016	100	−0.016	100
$\widehat{Y}_{ES}$			−.007	91.21	.008	88.68	.001	89.92
$\widehat{Y}_{opt_D}$			−.015	102.86	−.015	102.86	−.015	102.86

Note: m Denotes the size of primary stage units and $\bar{n}$ the average size over $B = 1,000$ simulations of final stage units. Linear, raking, and logit are used as calibration methods. Main variable: Plausible value in mathematics. School-auxiliary variables: Type of school and type of community. Student-auxiliary variables: gender, mother educational level, father educational level, and highest educational level. Probability proportional to size sampling is considered.

in order to compute these estimators. In addition, the estimators are defined under arbitrary sampling designs, in which any sampling design is used at the student level, given a number of schools drawn using any sampling design from the population.

The proposed method is similar to the Estevao–Särndal approach, and both methodologies use the same auxiliary information. For this reason, the Estevao–Särndal estimator is considered as a competitor estimator in the simulation studies. Nevertheless, there are differences; for example, the conditions for the calibration equations are less restrictive than those involved in the Estevao–Särndal method, which can imply a computational advantage. The aim of the proposed method is to obtain an estimator at the unit level,

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

Table 6. Values of RB% and RE% for the Various Estimators and the HBS Population.

	m	$\bar{n}$	Linear		Ranking		Logit	
			RB%	RE%	RB%	RE%	RB%	RE%
$\widehat{Y}_{\pi}$	25	48.058	.385	100	.385	100	.385	100
$\widehat{Y}_{ES}$			.148	100.86	−.009	96.09	−.026	97.10
$\widehat{Y}_{opt_D}$			.090	133.52	.035	132.71	.038	132.84
$\widehat{Y}_{\pi}$	50	96.576	−.134	100	−.134	100	−.134	100
$\widehat{Y}_{ES}$			.322	126.02	.163	125.20	.183	125.39
$\widehat{Y}_{opt_D}$			.265	139.46	.241	139.19	.243	139.27
$\widehat{Y}_{\pi}$	75	145.027	.251	100	.251	100	.251	100
$\widehat{Y}_{ES}$			.352	136.58	.250	133.92	.258	134.86
$\widehat{Y}_{opt_D}$			.318	146.04	.307	145.92	.308	145.99
$\widehat{Y}_{\pi}$	100	193.04	.072	100	.072	100	.072	100
$\widehat{Y}_{ES}$			.101	146.46	.010	145.36	.021	145.65
$\widehat{Y}_{opt_D}$			.098	147.43	.085	147.64	.086	147.62
$\widehat{Y}_{\pi}$	150	289.01	.102	100	.102	100	.102	100
$\widehat{Y}_{ES}$			.301	139.51	.231	140.03	.240	139.97
$\widehat{Y}_{opt_D}$			.175	146.31	.166	146.41	.167	146.40
$\widehat{Y}_{\pi}$	200	385.727	−.124	100	−.124	100	−.124	100
$\widehat{Y}_{ES}$			−.005	138.40	−.056	137.80	−.050	137.94
$\widehat{Y}_{opt_D}$			−.061	139.78	−.067	139.70	−.066	139.71
$\widehat{Y}_{\pi}$	250	481.883	−.085	100	−.085	100	−.085	100
$\widehat{Y}_{ES}$			.022	132.57	−.022	133.13	−.017	133.06
$\widehat{Y}_{opt_D}$			.089	136.55	.086	136.56	.086	136.56

Note: m Denotes the size of primary stage units and $\bar{n}$ the average size over $B = 1{,}000$ simulations of final stage units. Linear, raking, and logit are used as calibration methods. Main variable: Income. Household-auxiliary variable: Main breadwinner educational level. Person-auxiliary variables: gender, educational level, and age (as continuous variable). Probability proportional to size sampling is considered.

whereas the Estevao–Särndal estimator can be used to obtain estimators at the cluster level.

The estimators proposed are more efficient than standard estimator. This issue, according to our simulation study, is not numerically satisfied by the Estevao–Särndal estimator. The gain in efficiency of the proposed estimator over the Estevao–Särndal estimator is negligible for large sample sizes, but it can be larger for small sample sizes. This implies that the proposed method may be useful for users of education surveys, since such users are not usually large companies, and tend to use surveys with small sample sizes. It is interesting to note that in our simulation study the PISA database has been used as population.

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

Notes

Calibration estimators can be computed using the Sampling package (Tillé and Matei 2009).

`calib` computes the g weights of the calibration estimator (`g` $= w_k/d_k$ or $g_{Ii} = w_{Ii}/d_{Ii}$). It requires `Xs`$(= \mathbf{x}_{(u)k}$ or $\mathbf{x}_{(c)i})$ as the matrix of calibration variables, `d`$(= d_k$ or $d_{Ii})$ as the vector of initial weights, `total`$(= \sum_{k\in U} \mathbf{x}_{(u)k}$ or $\sum_{i\in U_I} \mathbf{x}_{(c)i})$ as the vector of population totals and method as the calibration `method` (linear, raking, or logit).

`calibev` computes the calibration estimator and its variance estimation using the residuals technique. The arguments are `Ys`$(= y_{(u)k})$, vector of interest variable; its size is n, the sample size, `Xs`, matrix of sample calibration variables, total, vector of population totals for calibration, `pikl`$(= \pi_{k\ell}$ or $\pi_{Iij})$, matrix of joint inclusion probabilities of the sample units, `d`, vector of initial weights of the sample units and `g`, vector of g weights. By default with=TRUE, the variance estimation takes into account the final weights `w=g*d`, otherwise `with=FALSE`, the initial weights d are taken into account.

The `R` code to compute the proposed estimators can be obtained from the authors on request.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This Research is partially supported by Ministerio de Educación y Ciencia (grant MTM2012-35650 and MTM2009-10055, Spain).

References

Copas, J. B. 1983. "Regression, Prediction and Shrinkage." *Journal of the Royal Statistical Society. Series B (Methodological)* 45:311-54.

Copas, J. B. 1993. "The Shrinkage of Point Scoring Methods." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 42:315-31.

Chambers, R. L. and R. Dunstan. 1986. "Estimating Distribution Function from Survey Data." *Biometrika* 73:597-604.

Deville, J. C. and C. E. Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87:376-82.

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

Deville, J. C., C. E. Särndal, and O. Sautory 1993. "Generalized Raking Procedures in Survey Sampling." *Journal of the American Statistical Association* 88:1013-20.

Estevao, V. M. and C. E. Särndal. 2006. "Survey Estimates by Calibration on Complex Auxiliary Information." *International Statistical Review* 74:127-47.

Gambino, J. G. and P. L. N. Silva. 2009. "Sampling and Estimation in Household Surveys in Handbook of Statistics." Pp. 407–439 in *Sample Surveys: Design, Methods and Applications* Vol. 29A, edited by D. Pfeffermann and C. R. Rao. North-Holland, The Netherlands: Elservier B. V.

Gershunskaya, J., J. Jiang, and P. Lahiri. 2009. "Resampling Methods in Surveys." in *Handbook of Statistics—Sample Surveys: Inference and Analysis* (Vol. 29B, pp. 123-151). edited by D. Pfeffermann and C. R. Rao. North-Holland, the Netherlands: Elservier B. V.

Gijbels, I. and N. Veraverbeke. 1988. "Weak Asymptotic Representations for Quantiles of the Product-limit Estimator." *Journal of Statistical Planning and Inference* 18:151-60.

INE. Microdata file Household budget survey (Base 2006) Retrieved from http://www.ine.es/en/prodyser/micro_epf2006_en.htm

Isaki, C. T. and W. A. Fuller. 1982. "Survey Design under the Regression Superpopulation Model." *Journal of the American Statistical Association* 77:89-96.

Jia, Y., L. Stokes, I. Harris, and Y. Wang. 2011. "Performance of Random Effects Model Estimators Under Complex Sampling Design." *Journal of Educational and Behavioral Statistics* 36:6–32.

Johnson, E. G. and K. F. Rust. 1992. "Population Inferences and Variance Estimation for NAEP Data." *Journal of Educational and Behavioral Statistics* 20:175-90.

Kirk M. Wolter 2007. *Introduction to Variance Estimation*, 2nd ed. NY, USA: Springer.

Krapavickaite, D. and A. Plikusas. 2005. "Estimation of a Ratio in the Finite Population." *Informatica* 16:347-64.

Kovacevic, M. S. 1997. "Calibration Estimation of Cumulative Distribution and Quantile Functions from Survey Data." *Proceedings of the Survey Methods Section, Statistical Society of Canada,* Meeting 1-7.

LeGuennec, J. and O. Sautory. 2002. CALMAR2: une nouvelle version de la macro CALMAR de redressement déchantillon par calage. Actes des Journées de Méthodologie, INSEE, Paris, France.

Lemaître, G. E. and J. Dufour. 1987. "An Integrated Method for Weighting Persons and Families." *Survey Methodology* 13:199-207.

Midzuno, H. 1952. "On the Sampling System with Probability Proportionate to Sum of Sizes." *Annals of the Institute of Statistical Mathematics* 3:99-107.

Pfeffermann, D. 1993. "The Role of Sampling Weights When Modeling Survey Data. " *International Statistical Review* 61:317-37.

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

Pfeffermann, D., C. J. Skinner, D. J. Holmes, H. Goldstein, and J. Rasbash. 1998. "Weighting for Unequal Selection Probabilities in Multilevel Models." *Journal of the Royal Statistical Society, Series B* 60:23-40.

PISA. Database 2006, Retrieved from http://pisa2006.acer.edu.au/

Plikusas, A. 2006. "Nonlinear Calibration." Pp. 52–57 in *Workshop on Survey Sampling Theory and Methods*. Central Statistical Bureau of Latvia. Ventspils, Latvia.

Randles, R. H.. 1982. "On the Asymptotic Normality of Statistics with Estimated Parameters." *Annals of Statistics* 10:462-74.

Rao, J. N. K. 2003. *Small Area Estimation*. Hoboken, New Jersey: John Wiley & Sons.

Rao, J. N. K., J. G. Kovar, and H. J. Mantel. 1990. "On Estimating Distribution Functions and Quantiles from Survey Data Using Auxiliary Information." *Biometrika* 77:365-75.

Rueda, C. and J. A. Menéndez. 2010. "The Selection of the Shrinkage Region in Small Area Estimation." *Advances in Soft Computing* 77:553-60.

Rueda, M., S. Martínez, H. Martínez, and A. Arcos. 2007. "Estimation of the Distribution Function with Calibration Methods." *Journal of Statistical Planning and Inference* 137:435-48.

Rust, K. F. and E. G. Johnson. 1992. "Sampling and Weighting in the National Assessment." *Journal of Educational and Behavioral Statistics* 20:111-29.

Särndal, C. E. 2007. "The Calibration Approach in Survey Theory and Practice." *Survey Methodology* 33:99-119.

Särndal, C. E. 1982. "Implications of Survey Design for Generalized Regression Estimation of Linear Functions." *Journal of Statistical Planning and Inference* 7:155-70.

Schäfer, J. and K. Strimmer. 2005. "A Shrinkage Approach to Large-scale Covariance Matrix Estimation and Implications for Functional Genomics." *Statistical Applications in Genetics and Molecular Biology* 4:1175-1189.

Silva, P. L. D. N. 2003. "Calibration Estimation: When and Why, How Much and How." Technical Report, University of Southampton, England.

Singh, S. 2001. "Generalized Calibration Approach for Estimating Variance in Survey Sampling." *Annals of the Institute of Statistical Mathematics* 53:404-17.

Singh, S. 2004. "Repair of Two-phase Calibration Methodology in Survey Sampling." SSC Conference at Montreal, Canada.

Singh, S. 2012. "On the Calibration of Design Weights Using a Displacement Function." *Metrika* 75:85-107. doi: 10.1007/s00184-010-0316-6.

Singh, S., S. Horn, S. Chowdhury, and F. Yu. 1999. "Calibration of the Estimators of Variance." *Australian and New Zealand Journal of Statistics* 41:199-212.

Thompson, J. R. 1968. "Some Shrinkage Technique for Estimating the Mean." *Journal of the American Statistical Association* 63:113-23.

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015

Tillé, Y. and A. Matei. 2009. Sampling: Survey Sampling. R package version 2.2. Retrieved from http://cran.r-project.org/src/contrib/Descriptions/sampling.html

Tillé, Y. 2002. "Unbiased Estimation by Calibration on Distribution in Simple Sampling Designs Without Replacement." *Survey Methodology* 28:77-85.

Wang, J. C. and J. Opsomer. 2011. "On the Asymptotic Normality and Variance Estimation of Non Differentiable Survey Estimators." *Biometrika* 98:91106

James, W. and Charles Stein. 1961. "Estimation with Quadratic Loss." Pp. 361–379 in *Proc. Fourth Berkeley Symp. on Math. Statist. and Prob*. Vol. 1. Univ. of Calif. Press.

Author Biographies

Antonio Arcos, PhD is a Professor at the Department of Statistics and Operational Research, University of Granada. He specializes in the design and analysis of complex surveys. His research has focused on using auxiliary information in surveys in order to improve the accuracy of survey estimates.

José M. Contreras, PhD is an Associate Professor of Mathematics Education at the University of Granada. His research interests include Statistic Education and Survey Estimates.

María M. Rueda, PhD is Full Professor at the Department of Statistics and Operational Research, University of Granada. Her professional interests include teaching and study of survey research methods, with particular emphasis on the use of auxiliary information in complex survey designs, methods for the analysis of survey data and survey nonresponse.

Downloaded from smr.sagepub.com at Biblioteca Universitaria de Granada on February 17, 2015