Catalogue no. 12-001-X

Survey Methodology

June 2010

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website at www.statcan.gc.ca, e-mail us at infostats@statcan.gc.ca, or telephone us, Monday to Friday from 8:30 a.m. to 4:30 p.m., at the following numbers:

Statistics Canada's National Contact Centre

Toll-free telephone (Canada and United States):

Inquiries line	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Fax line	1-877-287-4369

Local or international calls:

Inquiries line	1-613-951-8116
Fax line	1-613-951-0581

Depository Services Program

Inquiries line	1-800-635-7943
Fax line	1-800-565-7757

To access and order this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website at www.statcan.gc.ca and select "Publications."

This product, Catalogue no. 12-001-X, is also available as a standard printed publication at a price of CAN$30.00 per issue and CAN$58.00 for a one-year subscription.

The following additional shipping charges apply for delivery outside Canada:

	Single issue	Annual subscription
United States	CAN$6.00	CAN$12.00
Other countries	CAN$10.00	CAN$20.00

All prices exclude sales taxes.

The printed version of this publication can be ordered as follows:

- Telephone (Canada and United States) 1-800-267-6677
- Fax (Canada and United States) 1-877-287-4369
- E-mail infostats@statcan.gc.ca
- Mail Statistics Canada
Finance
R.H. Coats Bldg., 6th Floor
150 Tunney's Pasture Driveway
Ottawa, Ontario K1A 0T6
- In person from authorized agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "About us" > "Providing services to Canadians."

Statistics Canada

Business Survey Methods Division

Survey Methodology

June 2010

Published by authority of the Minister responsible for Statistics Canada

June 2010

Catalogue no. 12-001-XIE
ISSN 1492-0921

Catalogue no. 12-001-XPB
ISSN: 0714-0045

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-X au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is indexed in The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods. It is also covered by SCOPUS in the Elsevier Bibliographic Databases.

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (smj@statcan.gc.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca).

Subscription Rates

The price of printed versions of *Survey Methodology* (Catalogue No. 12-001-XPB) is CDN $58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN $12 ($6 × 2 issues); Other Countries, CDN $20 ($10 × 2 issues). A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec. Electronic versions are available on Statistics Canada's web site: www.statcan.gc.ca.

Survey Methodology

A Journal Published by Statistics Canada

Volume 36, Number 1, June 2010

Contents

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.
∞

Le papier utilisé dans la présente publication répond aux exigences minimales de l'"American National Standard for Information Sciences" – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.
∞

Survey Methodology, June 2010
Vol. 36, No. 1, p. 1
Statistics Canada, Catalogue No. 12-001-X

Cochran-Hansen Prize 2011

Competition for Young Survey Statisticians from Developing and Transitional Countries

In celebration of its 25th anniversary, the International Association of Survey Statisticians (IASS) established the Cochran-Hansen Prize to be awarded every two years to the best paper on survey research methods submitted by a young statistician from a developing or transitional country.

Participation in the competition for the Prize is open to nationals of developing or transitional countries who are living in such countries and who were born in 1971 or later.

Papers submitted must be unpublished original works. They may include materials from the participant's university thesis. They should be in either English or French. Papers for consideration should be submitted to the IASS Secretariat at the address below to arrive by December 29, 2010. Each submission should be accompanied by a cover letter that gives the participant's year of birth, nationality, and country of residence. The cover letter must also indicate if the work submitted is the result of a PhD thesis and, in the case of joint papers, the prize candidate must state clearly what his/her contribution to the paper is.

The papers submitted will be examined by the Cochran-Hansen Prize Committee appointed by the IASS. The decision of the Committee is final.

The winner of the Prize will be invited to present his/her paper at the 58th Session of the International Statistical Institute to be held in Dublin, Ireland, August 21-29, 2011, and the name of the winner will be announced at the ISI General Assembly in Dublin.

The author of the winning paper will receive the Cochran-Hansen Prize in the form of books and journal subscriptions to the value of about € 500, and will have reasonable travel and living expenses paid in order to present the paper at the ISI Session in Dublin.

For further information, please contact:

Madame Claude OLIVIER
IASS Secretariat
International Association of Survey Statisticians
CEFILINSEE, 3 rue de la Cité, 33500 Libourne, France
Tel: +33 5 57 55 56 17
Fax: +33 5 57 55 56 20
E-mail: Claude.olivier@insee.fr

Survey Methodology, June 2010
Vol. 36, No. 1, pp. 3-11
Statistics Canada, Catalogue No. 12-001-X

Fence method for nonparametric small area estimation

Jiming Jiang, Thuan Nguyen and J. Sunil Rao [1]

Abstract

This paper considers the problem of selecting nonparametric models for small area estimation, which recently have received much attention. We develop a procedure based on the idea of fence method (Jiang, Rao, Gu and Nguyen 2008) for selecting the mean function for the small areas from a class of approximating splines. Simulation results show impressive performance of the new procedure even when the number of small areas is fairly small. The method is applied to a hospital graft failure dataset for selecting a nonparametric Fay-Herriot type model.

Key Words: Fay-Herriot Model; Fence method; Nonparametric model selection; Penalized spline; Small area estimation.

1. Introduction

Small area estimation (SAE) has received increasing attention in recent literature. Here the term small area typically refers to a population for which reliable statistics of interest cannot be produced due to certain limitations of the available data. Examples of small areas include a geographical region (*e.g.*, a state, county, municipality, *etc.*), a demographic group (*e.g.*, a specific age × sex × race group), a demographic group within a geographic region, *etc*. In absence of adequate direct samples from the small areas, methods have been developed in order to "borrow strength". Statistical models, especially mixed effects models, have played important roles in SAE. See Rao (2003) for a comprehensive account of various methods used in SAE.

While there is extensive literature on inference about small areas using mixed effects models, including estimation of small area means which is a problem of mixed model prediction, estimation of the mean squared error (MSE) of the empirical best linear unbiased predictor (EBLUP; see Rao 2003), and prediction intervals (*e.g.*, Chatterjee, Lahiri and Li 2007), model selection in SAE has received much less attention. However, the importance of model selection in SAE has been noted by prominent researchers in this field (*e.g.*, Battese, Harter and Fuller 1988, Ghosh and Rao 1994). Datta and Lahiri (2001) discussed a model selection method based on computation of the frequentist's Bayes factor in choosing between a fixed effects model and a random effects model. They focused on the following one-way balanced random effects model for the sake of simplicity: $y_{ij} = \mu + u_i + e_{ij}$, $i = 1, ..., m$, $j = 1, ..., k$, where the u_i's and e_{ij}'s are normally distributed with mean zero and variances σ_u^2 and σ_e^2, respectively. As noted by the authors, the choice between a fixed effects model and a random effects one in this case is equivalent to testing the following one-sided hypothesis $H_0: \sigma_u^2 = 0$ *vs* $H_1: \sigma_u^2 > 0$. Note that, however, not all model selection problems can be formulated as hypothesis testing. Fabrizi and Lahiri (2004) developed a robust model selection method in the context of complex surveys. Meza and Lahiri (2005) demonstrated the limitations of Mallows' C_p statistic in selecting the fixed covariates in a nested error regression model (Battese, Harter and Fuller 1988), defined as $y_{ij} = x_{ij}'\beta + u_i + e_{ij}$, $i = 1, ..., m$, $j = 1, ..., n_i$, where y_{ij} is the observation, x_{ij} is a vector of fixed covariates, β is a vector of unknown regression coefficients, and u_i's and e_{ij}'s are the same as in the model above considered by Datta and Lahiri (2001). Simulation studies carried out by Meza and Lahiri (2005) showed that the C_p method without modification does not work well in the current mixed model setting when the variance σ_u^2 is large; on the other hand, a modified C_p criterion developed by these latter authors by adjusting the intra-cluster correlations performs similarly as the C_p in regression settings. It should be pointed out that all these studies are limited to linear mixed models, while model selection in SAE in a generalized linear mixed model (GLMM) setting has never been seriously addressed.

Recently, Jiang *et al.* (2008) developed a new strategy for model selection, called *fence methods*. The authors noted a number of limitations of the traditional model selection strategies when applied to mixed model situations. For example, the BIC procedure (Schwarz 1978) relies on the effective sample size which is unclear in typical situations of SAE. To illustrate this, consider the nested error regression model introduced above. Clearly, the effective sample size is not the total number of observations $n = \sum_{i=1}^m n_i$, neither is proportional to m, the number of small areas unless all the n_i are equal and fixed. The fence methods avoid such

1. Jiming Jiang, University of California, Davis. E-mail: jiang@wald.ucdavis.edu; Thuan Nguyen, Oregon Health and Science University; J. Sunil Rao, Case Western Reserve University.

limitations, and therefore are suitable to mixed model selection problems, including linear mixed models and GLMMs. The basic idea of fence is to build a statistical fence to isolate a subgroup of what are known as the correct models. Once the fence is constructed, the optimal model is selected from those within the fence according to a criterion which can incorporate quantities of practical interest. More details about the fence methods are given below.

The focus of this paper is nonparametric models for SAE. These models have received much recent attention. In particular, Opsomer, Breidt, Claeskens, Kauermann and Ranalli (2007) proposed a spline-based nonparametric model for SAE. The idea is to approximate an unknown nonparametric small-area mean function by a penalized spline (P-spline). The authors then used a connection between P-splines and linear mixed models (Wand 2003) to formulate the approximating model as a linear mixed model, where the coefficients of the splines are treated as random effects. Consider, for simplicity, the case of univariate covariate. Then, a P-spline can be expressed as

$$\tilde{f}(x) = \beta_0 + \beta_1 x + ... + \beta_p x^p + \gamma_1 (x - \kappa_1)_+^p + ... + \gamma_q (x - \kappa_q)_+^p, \quad (1)$$

where p is the degree of the spline, q is the number of knots, $\kappa_j, 1 \le j \le q$ are the knots, and $x_+ = x1_{(x>0)}$. Clearly, a P-spline is characterized by p, q, and also the location of the knots. Note that, however, given p, q, the location of the knots can be selected by the space-filling algorithm implemented in R [*cover.design*()]. But the question how to choose p and q remains. The general "rule of thumb" is that p is typically between 1 and 3, and q proportional to the sample size, n, with 4 or 5 observations per knot (Ruppert, Wand and Carroll 2003). But there may still be a lot of choices given the rule of thumb. For example, if $n = 200$, the possible choices for q range from 40 to 50, which, combined with the range of 1 to 3 for p, gives a total of 33 choices for the P-spline. Our new adaptive fence method offers a data-driven approach for choosing p and q for the spline-based SAE model.

The rest of the paper is organized as follows. The fence methods are described in section 2. In section 3 we develop an adaptive fence procedure for the nonparametric model selection problem. In section 4 we demonstrate the finite sample performance of the new procedure with a series of simulation studies. In section 5 we consider a real-life data example involving a dataset from a medical survey which has been used for fitting a Fay-Herriot model (Fay and Herriot 1979). Some technical results are deferred to the appendix.

2. Fence methods

As mentioned, the basic idea of fence is to construct a statistical fence and then select an optimal model from those within the fence according to certain criterion of optimality, such as model simplicity. Let $Q_M = Q_M(y, \theta_M)$ be a measure of lack-of-fit, where y represents the vector of observations, M indicates a candidate model, and θ_M denotes the vector of parameters under M. Here by lack-of-fit we mean that Q_M satisfies the basic requirement that $\mathrm{E}(Q_M)$ is minimized when M is a true model, and θ_M the true parameter vector under M. Then, a candidate model M is in the fence if

$$\hat{Q}_M \le \hat{Q}_{\tilde{M}} + c_n \hat{\sigma}_{M,\tilde{M}}, \quad (2)$$

where $\hat{Q}_M = \inf_{\theta_M \in \Theta_M} Q_M$, Θ_M being the parameter space under M, $\tilde{M}$ is a model that minimizes $\hat{Q}_M$ among $M \in \mathcal{M}$, the set of candidate models, and $\hat{\sigma}_{M,\tilde{M}}$ is an estimate of the standard deviation of $\hat{Q}_M - \hat{Q}_{\tilde{M}}$. The constant c_n on the right side of (2) can be chosen as a fixed number (*e.g.*, $c_n = 1$) or adaptively (see below).

The calculation of $\hat{Q}_M$ is usually straightforward. For example, in many cases Q_M can be chosen as the negative log-likelihood, or residual sum of squares. On the other hand, the computation of $\hat{\sigma}_{M,\tilde{M}}$ can be quite challenging. Sometimes, even if an expression can be obtained for $\hat{\sigma}_{M,\tilde{M}}$, its accuracy as an estimate of the standard deviation cannot be guaranteed in a finite sample situation. Jiang, Nguyen and Rao (2009) simplified an adaptive fence procedure proposed by Jiang *et al.* (2008). For simplicity, we assume that $\mathcal{M}$ contains a full model, M_{f}, of which each candidate model is a submodel. It follows that $\tilde{M} = M_{\mathrm{f}}$. In the simplified adaptive procedure, the fence inequality (2) is replaced by

$$\hat{Q}_M - \hat{Q}_{M_{\mathrm{f}}} \le c_n, \quad (3)$$

where c_n is chosen adaptively as follows. For each $M \in \mathcal{M}$, let $p^*(M) = \mathrm{P}^*\{M_0(c) = M\}$ be the empirical probability of selection for M, where $M_0(c)$ denotes the model selected by the fence procedure based on (3) with $c_n = c$, and P^* is obtained by bootstrapping under M_{f}. For example, under a parametric model one can estimate the model parameters under M_{f} and then use a parametric bootstrap to draw samples under M_{f}. Suppose that B samples are drawn, then $p^*(M)$ is simply the sample proportion (out of a total of B samples) that M is selected by the fence procedure based on (3) with the given c_n. Let $p^* = \max_{M \in \mathcal{M}} p^*(M)$. Note that p^* depends on c_n. Let c_n^* be the c_n that maximizes p^* and this is our choice. Jiang *et al.* (2008) offers the following explanation of the motivation behind adaptive fence. Suppose that there is a true model among the candidate models, then, the optimal model is the one from which the data is generated, and

therefore should be the most likely given the data. Thus, given c_n, one is looking for the model (using the fence procedure) that is most supported by the data or, in other words, one that has the highest (posterior) probability. The latter is estimated by bootstrapping. Note that although the bootstrap samples are generated under M_{f}, they are almost the same as those generated under the optimal model. This is because the estimates corresponding to the zero parameters are expected to be close to zero, provided that the parameter estimators under M_{f} are consistent. One then pulls off the c_n that maximizes the (posterior) probability and this is the optimal choice.

There are two extreme cases corresponding to $c_n = 0$ and $c_n = \infty$ (*i.e.*, very large). Note that if $c_n = 0$, then $p^* = 1$. This is because when $c_n = 0$ the procedure always chooses M_{f}. Similarly, if there is a unique simplest model (*e.g.*, model with minimum dimension), say, M_*, then $p^* = 1$ for very large c_n. This is because, when c_n is large enough, all models are in the fence, hence the procedure always chooses M_*, if simplicity is used as the criterion of optimality for selecting the model within the fence. These two extreme cases are handled carefully in Jiang *et al.* (2008) and Jiang *et al.* (2009). However, as noted by Jiang *et al.* (2008), the procedures to handle the extreme cases, namely, the screen tests and baseline adjustment/threshhold checking, are rarely needed in practice. For example, in most applications there are a (large) number of candidate variables, and it is believed that only a (small) subset of them are important. This means that the optimal model is neither M_* nor M_{f}. Therefore, there is no need to worry about the extreme cases, and the procedures to handle these cases can be skipped. In most applications a plot of p^* against c_n is W-shaped with the peak in the middle corresponding to c_n^*.

The left plot of Figure 2 provides an illustration. This is a plot of p^* against c_n for the example discussed in section 5. The plot shows the typical "W" shape, as described, and the peak in the middle corresponds to where the optimal c_n, *i.e.*, c_n^* is.

Jiang *et al.* (2009) established consistency of the simplified adaptive fence and studied its finite sample performance.

3. Nonparametric SAE model selection

For the simplicity of illustration we consider the following SAE model:

$$y_i = f(X_i) + B_i u_i + e_i, \quad i = 1, \ldots, m, \tag{4}$$

where y_i is an $n_i \times 1$ vector representing the observations from the i^{th} small area; $f(X_i) = [f(x_{ij})]_{1 \le j \le n_i}$ with $f(x)$ being an unknown (smooth) function; B_i is an $n_i \times b$ known matrix; u_i is a $b \times 1$ vector of small-area specific random effects; and e_i is an $n_i \times 1$ vector of sampling errors. It is assumed that u_i, e_i, $i = 1, \ldots, m$ are independent with $u_i \sim N(0, G_i)$, $G_i = G_i(\theta)$, and $e_i \sim N(0, R_i)$, $R_i = R_i(\theta)$, θ being an unknown vector of variance components. Note that, besides $f(X_i)$, the model is the same as the standard "longitudinal" linear mixed model (*e.g.*, Laird and Ware 1982, Datta and Lahiri 2000).

The approximating spline model is given by replacing $f(x)$ by $\tilde{f}(x)$ in (1), where the coefficients β's and γ's are estimated by penalized least squares, *i.e.*, by

$$\text{minimizing} \quad |y - X\beta - Z\gamma|^2 + \lambda |\gamma|^2, \tag{5}$$

where $y = (y_i)_{1 \le i \le m}$, the $(i, j)^{\text{th}}$ row of X is $(1, x_{ij}, \ldots, x_{ij}^p)$, the $(i, j)^{\text{th}}$ row of Z is $[(x_{ij} - \kappa_1)_+^p, \ldots, (x_{ij} - \kappa_q)_+^p]$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, and λ is a penalty, or smoothing, parameter. To determine λ, Wand (2003) used the following interesting connection to a linear mixed model. To illustrate the idea, let us consider a simple case in which $B_i = 0$ (*i.e.*, there is no small-area random effects), and the components of e_i are independent and distributed as $N(0, \tau^2)$. If the γ's are treated as random effects which are independent and distributed as $N(0, \sigma^2)$, then the solution to (5) are the same as the best linear unbiased estimator (BLUE) for β, and the best linear unbiased predictor (BLUP) for γ, if λ is identical to the ratio τ^2/σ^2. Thus, the value of λ may be estimated by the maximum likelihood (ML), or restricted maximum likelihood (REML) estimators of σ^2 and τ^2 (*e.g.*, Jiang 2007). However, there has been study suggesting that this approach is biased towards undersmoothing (Kauermann 2005). Consider, for example, a special case in which $f(x)$ is, in fact, the quadratic spline with two knots given by (10). (Note that this function is smooth in that it has a continuous derivative.) It is clear that, in this case, the best approximating spline should be $f(x)$ itself with only two knots, *i.e.*, $q = 2$ (of course, one could use a spline with many knots to "approximate" the two-knot quadratic spline, but that would seem very inefficient in this case). However, if one uses the above linear mixed model connection, the ML (or REML) estimator of σ^2 is consistent only if $q \to \infty$ (*i.e.*, the number of appearances of the spline random effects goes to infinity). The seeming inconsistency has two worrisome consequences: (i) the meaning of λ may be conceptually difficult to interpret; (ii) the behavior of the estimator of λ may be unpredictable.

The fence method offers a natural approach to choosing the degree of the spline, p, the number of knots, q, and the smoothing parameter, λ at the same time. Note, however, a major difference from the situations considered in Jiang *et al.* (2008) and Jiang *et al.* (2009) in that the true underlying model is not among the class of candidate models, *i.e.*, the approximating splines (1). Furthermore, the

role of λ in the model should be made clear: λ controls the degree of smoothness of the underlying model. A natural measure of lack-of-fit is $Q_M = |y - X\beta - Z\gamma|^2$. However, $\hat{Q}_M$ is not obtained by minimizing Q_M over β and γ without constraint. Instead, we have $\hat{Q}_M = |y - X\hat{\beta} - Z\hat{\gamma}|^2$, where $\hat{\beta}$ and $\hat{\gamma}$ are the solution to (5), and hence depends on λ. The optimal λ is to be selected by the fence method, together with p and q, as described below.

Another difference is that there may not be a full model among the candidate models. Therefore, the fence inequality (3) is replaced by the following:

$$\hat{Q}_M - \hat{Q}_{\tilde{M}} \le c_n, \tag{6}$$

where $\tilde{M}$ is the candidate model that has the minimum $\hat{Q}_M$. We use the following criterion of optimality within the fence which combines model simplicity and smoothness. For the models within the fence, choose the one with the smallest q; if there are more than one such models, choose the model with the smallest p. This gives the best choice of p and q. Once p, q are chosen, we choose the model *within the fence* with the largest λ. Once again, note that λ is part of the model M that is selected (or "estimated") by the fence method. The tuning constant c_n is chosen adaptively using the simplified adaptive procedure of Jiang *et al.* (2009), where parametric bootstrap is used for computing p^* (see section 2).

The following theorem is proved in Appendix. For simplicity, assume that the matrix $W = (X\ Z)$ is of full rank. Let $P_{W^\perp} = I_n - P_W$, where $n = \sum_{i=1}^m n_i$ and $P_W = W(W'W)^{-1}W'$.

Theorem. Computationally, the above fence procedure is equivalent to the following: (i) first use the (adaptive) fence to select p and q using (6) with $\lambda = 0$ and $\hat{Q}_M = y' P_{W^\perp} y$ (see Lemma below), and same criterion as above for choosing p, q within the fence; (ii) let M_0^* denotes the model corresponding to the selected p and q, find the maximum λ such that

$$\hat{Q}_{M_0^*,\lambda} - \hat{Q}_{\tilde{M}} \le c_n^*, \tag{7}$$

where for any model M with the corresponding X and Z, we have

$$\begin{aligned}
\hat{Q}_{M,\lambda} &= |y - X\hat{\beta}_\lambda - Z\hat{\gamma}_\lambda|^2,\\
\hat{\beta}_\lambda &= (X'V_\lambda^{-1}X)^{-1}X'V_\lambda^{-1}y,\\
\hat{\gamma}_\lambda &= \lambda^{-1}(I_q + \lambda^{-1}Z'Z)^{-1}Z'(y - X\hat{\beta}_\lambda),\\
X'V_\lambda^{-1}X &= X'X - \lambda^{-1}X'Z(I_q + \lambda^{-1}Z'Z)^{-1}Z'X,\\
X'V_\lambda^{-1}y &= X'y - \lambda^{-1}X'Z(I_q + \lambda^{-1}Z'Z)^{-1}Z'y,
\end{aligned}$$

and c_n^* is chosen by the adaptive fence procedure described in section 2 (V_λ is defined below but not directly needed here for the computation because of the last two equations).

Note that in step (i) of the Theorem one does not need to deal with λ. The motivation for (7) is that this inequality is satisfied when $\lambda = 0$, so one would like to see how far λ can go. In fact, the maximum λ is a solution to the equation $\hat{Q}_{M_0^*,\lambda} - \hat{Q}_{\tilde{M}} = c_n^*$. The purpose of the last two equations is to avoid direct inversion of $V_\lambda = I_n + \lambda^{-1}ZZ'$, whose dimension is equal to n, the total sample size. Note that V_λ does not have a block diagonal structure because of ZZ', so if n is large direct inversion of V_λ may be computationally burdensome.

The proof of the Theorem requires the following lemma, whose proof is given in Appendix.

Lemma. For any M and y, $\hat{Q}_{M,\lambda}$ is an increasing function of λ with $\inf_{\lambda>0} \hat{Q}_{M,\lambda} = \hat{Q}_M$.

4. Simulations

We consider an extension of the Fay-Herriot model (Fay and Herriot 1979) in a nonparametric setting. The model can be expressed as

$$y_i = f(x_i) + v_i + e_i,\ i = 1, \ldots, m, \tag{8}$$

where $v_i, e_i, i = 1, \ldots, m$ are independent such that $v_i \sim N(0, A)$, $e_i \sim N(0, D_i)$, where A is unknown but the sampling variance D_i is assumed known. The main difference from the traditional Fay-Herriot model is $f(x_i)$, where $f(x)$ is an unknown smooth function.

For simplicity we assume $D_i = D$, $1 \le i \le m$. Then, the model can be expressed as

$$y_i = f(x_i) + \varepsilon_i,\ i = 1, \ldots, m, \tag{9}$$

where $\varepsilon_i \sim N(0, \sigma^2)$ with $\sigma^2 = A + D$, which is unknown. Thus, the model is the same as the nonparametric regression model.

We consider three different cases that cover various situations and aspects. In the first case, Case 1, the true underlying function is a linear function, $f(x) = 1 - x$, $0 \le x \le 1$, hence the model reduces to the traditional Fay-Herriot model. The goal is to find out if fence can validate the traditional Fay-Herriot model in the case that it is valid. In the second case, Case 2, the true underlying function is a quadratic spline with two knots, given by

$$f(x) = 1 - x + x^2 - 2(x-1)_+^2 + 2(x-2)_+^2,\ 0 \le x \le 3 \tag{10}$$

(the shape is half circle between 0 and 1 facing up, half circle between 1 and 2 facing down, and half circle between 2 and 3 facing up). Note that this function is smooth in that it has a continuous derivative. Here we intend to investigate whether the fence can identify the true underlying function in the "perfect" situation, *i.e.*, when $f(x)$ itself is a spline. The last case, Case 3, is perhaps the most practical situation,

in which no spline can provide a perfect approximation to $f(x)$. In other words, the true underlying function is not among the candidates. In this case $f(x)$ is chosen as $0.5\sin(2\pi x),\ 0 \le x \le 1,$ which is one of the functions considered by Kauermann (2005).

We consider situations of small or medium sample size, namely, $m = 10, 15$ or 20 for Case I, $m = 30$, 40 or 50 for Case 2, and $m = 10, 30$ or 50 for Case 3. The covariate x_i are generated from the Uniform$[0, 1]$ distribution in Case 1, and from Uniform$[0, 3]$ in Case 2; then fixed throughout the simulations. Following Kauermann (2005), we let x_i be the equidistant points in Case 3. The error standard deviation σ in (9) is chosen as 0.2 in Case 1 and Case 2. This value is chosen such that the signal standard deviation in each case is about the same as the error standard deviation. As for Case 3, we consider three different values for $\sigma, 0.2, 0.5$ and 1.0. These values are also of the same order as the signal standard deviation in this case.

The candidate approximating splines for Case 1 and Case 2 are the following: $p = 0, 1, 2, 3,\ q = 0$ and $p = 1, 2, 3,\ q = 2, 5$ (so there are a total of 10 candidates). As for Case 3, following Kauermann (2005), we consider only linear splines (*i.e.*, $p = 1$); furthermore, we consider the number of knots in the range of the "rule of thumb" (*i.e.*, roughly 4 or 5 observations per knot; see section 1), plus the intercept model $(p = q = 0)$ and the linear model $(p = 1,\ q = 0)$. Thus, for $m = 10,\ q = 0,\ 2,\ 3;$ for $m = 30,\ q = 0, 6, 7, 8;$ and for $m = 50,\ q = 0, 10, 11, 12, 13.$

Table 1 shows the results based on 100 simulations under Case 1 and Case 2. As in Jiang *et al.* (2009), we consider both the highest peak, that is, choosing c_n with the highest p^*, and 95% lower bound (L.B.), that is, choosing a smaller c_n corresponding to a peak of p^* in order to be conservative, if the corresponding p^* is greater than the 95% lower bound of the p^* for any larger c_n that corresponds to a peak of p^*. It is seen that performance of the adaptive fence is satisfactory even with the small sample size. Also, it appears that the confidence lower bound method works better in smaller sample, but makes almost no difference in larger sample. These are consistent with the findings of Jiang *et al.* (2009).

Table 1
Nonparametric model selection - Case 1 and Case 2. Reported are empirical probabilities, in terms of percentage, based on 100 simulations that the optimal model is selected

	Case 1			Case 2		
Sample size	$m = 10$	$m = 15$	$m = 20$	$m = 30$	$m = 40$	$m = 50$
Highest Peak	62	91	97	71	83	97
Confidence L.B.	73	90	97	73	80	96

Table 2 shows the results for Case 3. Note that, unlike Case 1 and Case 2, here there is no optimal model (an optimal model must be a true model according to our definition). So, instead of giving the empirical probabilities of selecting the optimal model, we give the empirical distribution of the selected models in each case. It is apparent that, as σ increases, the distribution of the models selected becomes more spread out. A reverse pattern is observed as m increases. The confidence lower bound method appears to perform better in picking up a model with splines. Within the models with splines, fence seems to overwhelmingly prefer fewer knots than more knots.

Table 2
Nonparametric model selection - Case 3. Reported are empirical distributions, in terms of percentage, of the selected models

	Sample Size **# of Knots**	$m = 10$ **0, 2, 3**		$m = 30$ **0, 6, 7, 8**		$m = 50$ **0, 10, 11, 12, 13**	
		(p, q)	%	(p, q)	%	(p, q)	%
$\sigma = 0.2$	Highest Peak	(0, 0)	1	(1, 0)	9	(1, 10)	100
		(1, 0)	31	(1, 6)	91		
		(1, 2)	68				
	Confidence L.B.	(1, 0)	24	(1, 0)	9	(1, 10)	100
		(1, 2)	76	(1, 6)	91		
$\sigma = 0.5$	Highest Peak	(0, 0)	14	(1, 0)	21	(1, 0)	13
		(1, 0)	27	(1, 6)	77	(1, 10)	84
		(1, 2)	56	(1, 7)	2	(1, 11)	2
		(1, 3)	3			(1, 12)	1
	Confidence L.B.	(0, 0)	8	(1, 0)	8	(1, 0)	2
		(1, 0)	23	(1, 6)	89	(1, 10)	94
		(1, 2)	65	(1, 7)	3	(1, 11)	2
		(1, 3)	4			(1, 12)	2
$\sigma = 1$	Highest Peak	(0, 0)	27	(0, 0)	15	(0, 0)	10
		(1, 0)	20	(1, 0)	18	(1, 0)	26
		(1, 2)	49	(1, 6)	63	(1, 10)	60
		(1, 3)	4	(1, 7)	4	(1, 11)	2
						(1, 12)	2
	Confidence L.B.	(0, 0)	20	(0, 0)	1	(0, 0)	2
		(1, 0)	13	(1, 0)	13	(1, 0)	13
		(1, 2)	59	(1, 6)	82	(1, 10)	80
		(1, 3)	8	(1, 7)	4	(1, 11)	2
						(1, 12)	3

Note that the fence procedure allows us to choose not only p and q but also λ (see section 3). In each simulation we compute $\hat{\beta} = \hat{\beta}_\lambda$ and $\hat{\gamma} = \hat{\gamma}_\lambda$, given below (7), based on the λ chosen by the adaptive fence. The fitted values are calculated by (1) with β and γ replaced by $\hat{\beta}$ and $\hat{\gamma}$, respectively. We then average the fitted values over the 100 simulations. Figure 1 shows the average fitted values for the three cases $(m = 10, 30, 50)$ with $\sigma = 0.2$ under Case 3. The true underlying function values, $f(x_i) = 0.5\sin(2\pi x_i)$, $i = 1, \ldots, m$ are also plotted for comparison.

5. A real-life data example

We consider a dataset from Morris and Christiansen (1995) involving 23 hospitals (out of a total of 219 hospitals) that had at least 50 kidney transplants during a 27 month period (Table 3). The y_i's are graft failure rates for kidney transplant operations, that is, y_i = number of graft failures $/n_i$, where n_i is the number of kidney transplants at hospital i during the period of interest. The variance for graft failure rate, D_i, is approximated by $(0.2)(0.8)/n_i$, where 0.2 is the observed failure rate for all hospitals. Thus, D_i is assumed known. In addition, a severity index x_i is available for each hospital, which is the average fraction of females, blacks, children and extremely ill kidney recipients at hospital i. The severity index is considered as a covariate.

Table 3
Hospital data from Morris and Christiansen (1995)

Area	y_i	x_i	$\sqrt{D_i}$
1	0.302	0.112	0.055
2	0.140	0.206	0.053
3	0.203	0.104	0.052
4	0.333	0.168	0.052
5	0.347	0.337	0.047
6	0.216	0.169	0.046
7	0.156	0.211	0.046
8	0.143	0.195	0.046
9	0.220	0.221	0.044
10	0.205	0.077	0.044
11	0.209	0.195	0.042
12	0.266	0.185	0.041
13	0.240	0.202	0.041
14	0.262	0.108	0.036
15	0.144	0.204	0.036
16	0.116	0.072	0.035
17	0.201	0.142	0.033
18	0.212	0.136	0.032
19	0.189	0.172	0.031
20	0.212	0.202	0.029
21	0.166	0.087	0.029
22	0.173	0.177	0.027
23	0.165	0.072	0.025

Ganesh (2009) proposed a Fay-Herriot model for the graft failure rates. as follows: $y_i = \beta_0 + \beta_1 x_i + v_i + e_i$, where the v_i's are hospital-specific random effects and e_i's are sampling errors. It is assumed that v_i, e_i are independent with $v_i \sim N(0, A)$ and $e_i \sim N(0, D_i)$. Here the variance A is unknown. Based on the model Ganesh obtained credible intervals for selected contrasts. However, inspections of the raw data suggest some nonlinear trends, which raises the question on whether the fixed effects part of the model can be made more flexible in its functional form.

To answer this question, we consider the Fay-Herriot model as a special member of a class of approximating spline models discussed in section 3. More specifically, we assume

$$y_i = f(x_i) + v_i + e_i, \quad i = 1, \ldots, m, \tag{11}$$

where $f(x)$ is an unknown smooth function and everything else are the same as in the Fay-Herriot model. We then consider the following class of approximating spline models:

$$\hat{f}(x) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \gamma_1 (x - \kappa_1)_+^p + \ldots + \gamma_q (x - \kappa_q)_+^p \tag{12}$$

with $p = 0, 1, 2, 3$ and $q = 0, 1, \ldots, 6$ ($p = 0$ is only for $q = 0$). Here the upper bound 6 is chosen according to the "rule-of-thumb" (because $m = 23$, so $m/4 = 5.75$). Note that the Fay-Herriot model corresponds to the case $p = 1$ and $q = 0$. The question is then to find the optimal model, in terms of p and q, from this class.

We apply the adaptive fence method described in section 3 to this case. Here to obtain the bootstrap samples needed for obtaining c_n^*, we first compute the ML estimator under the model $\tilde{M}$, which minimizes $\hat{Q}_M = y' P_{W^\perp} y$ among the candidate models [*i.e.*, (12); see Theorem in section 3], then draw parametric bootstrap samples under model $\tilde{M}$ with the ML estimators treated as the true parameters. This is reasonable because $\tilde{M}$ is the best approximating model in terms of the fit, even though under model (11) there may not be a true model among the candidate models. The bootstrap sample size is chosen as 100.

The fence method selects the model $p = 3$ and $q = 0$, that is, a cubic function with no knots, as the optimal model. To make sure that the bootstrap sample size $B = 100$ is adequate, we repeated the analysis 100 times, each time using different bootstrap samples (recall in the adaptive fence one needs to draw bootstrap samples in order to determine c_n^*, so the question is whether different bootstrap samples lead to different results of model selection). All results led to the same model: a cubic function with no knots (even though the bootstrap-derived intermediate quantities, such as p^* and c_n^*, varied across bootstraps). We also ran the data analysis using $B = 1{,}000$, and selected model remained the same. Thus, it appears that the bootstrap sample size $B = 100$ is adequate. The left figure of Figure 2 shows the plot of p^* against c_n in the adaptive fence model selection.

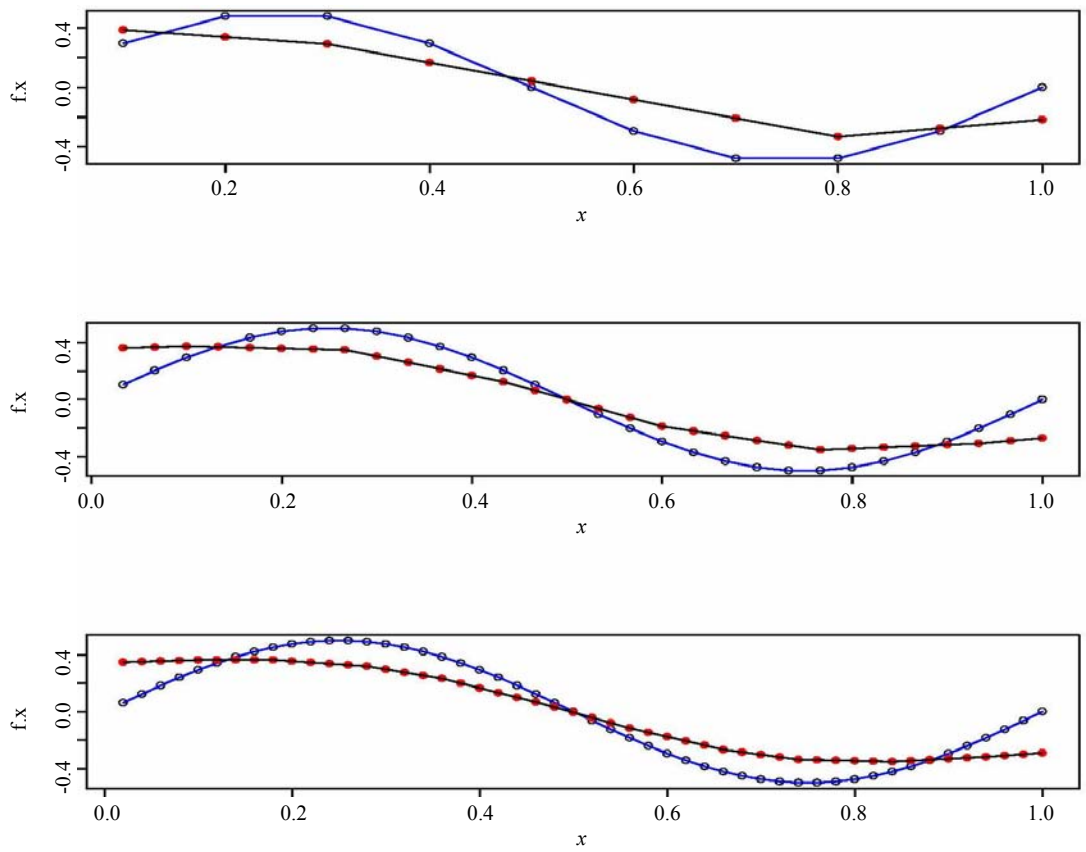

Figure 1 Case 3 Simulation. Top figure: Average fitted values for $m = 10$. Middle figure: Average fitted values for $m = 30$. Bottom figure: Average fitted values for $m = 50$. In all cases, the dots represent the fitted values, while the circles correspond to the true underlying function

A few comparisons are always helpful. Our first comparison is to fence itself but with a more restricted space of candidate models. More specifically, we consider (12) with the restriction to linear splines only, *i.e.*, $p = 1$, and knots in the range of the "rule of thumb", *i.e.*, $q = 4, 5, 6$, plus the intercept model $(p = q = 0)$ and the linear model $(p = 1,\ q = 0)$. In this case, the fence method selected a linear spline with four knots (*i.e.*, $p = 1,\ q = 4$) as the optimal model. The value of λ corresponding to this model is approximately equal to 0.001. The plot of p^* against c_n for this model selection is very similar to the left figure of Figure 2, and therefore omitted. In addition, the right figure of Figure 2 shows the fitted values and curves under the two models selected by the fence from within the different model spaces as well as the original data points.

A further comparison can be made by treating (11) as a generalized additive model (GAM) with heteroscedastic errors. A weighted fit can be obtained with the amount of smoothing optimized by using a generalized cross-validation (GCV) criterion. Here the weights used are $w_i = 1/(A + D_i)$ where the maximum likelihood estimate for A is used as a plug-in estimate. Recall that the D_i's are known. This fitted function is also overlayed in the right figure of Figure 2. Notice how closely this fitted function resembles the restricted space fence fit.

To expand the class of models under consideration by GCV-based smoothing, we used the BRUTO procedure (Hastie and Tibshirani 1990) which augments the class of models to look at a null fit and a linear fit for the spline function; and embeds the resulting model selection (*i.e.*, null, linear or smooth fits) into a weighted backfitting algorithm using GCV for computational efficiency. Interestingly here, BRUTO finds simply an overall linear fit for the fixed effects functional form. While certainly an interesting comparison, BRUTO's theoretical properties for models like (11) have not really been studied in depth.

Finally, as mentioned in section 3, by using the connection between P-spline and linear mixed model one can formulate (12) as a linear mixed model, where the spline coefficients are treated as random effects. The problem then becomes a (parametric) mixed model selection problem, hence the method of Jiang *et al.* (2009) can be applied. In fact, this was our initial approach to this dataset, and the model we found was the same as the one by BRUTO. However, we have some reservation about this approach, as explained in section 3.

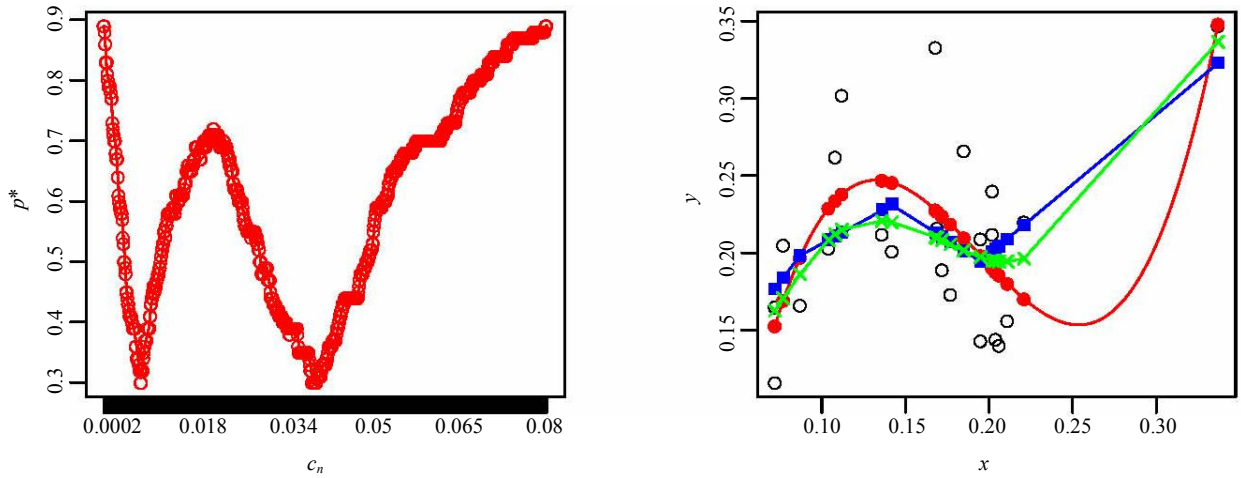

Figure 2 Left: A plot of p^* against c_n from the search over the full model space. Right: The raw data and the fitted values and curves; dots and their curve correspond to the cubic function resulted from the full model search; squares and their lines correspond to the linear spline with 4 knots resulted from the restricted model search; green X's and their lines represent the GAM fits

6. Concluding remarks

Although the focus of the current paper is nonparametric SAE model selection, our method may be applicable to spline-based mixed effects model selection problems in other areas, for example, in the analysis of longitudinal data (*e.g.*, Wang 2005).

In the case where a true model exists among the candidate models, such as Cases 1 and 2 in section 4, consistency of the proposed fence model selection method can be established in the same way as in Section 3 of Jiang *et al.* (2009) (although the result of the latter paper does not directly apply). However, practically, the situation that non-parametric modeling is most useful is when a true model does not exist, or is not among the candidates, such as Case 3 in section 4. In this case, no result of consistency can be proved, of course. It remains unclear what is a desirable asymptotic behavior to study in the latter case.

Acknowledgements

Jiming Jiang is partially supported by NSF grants DMS - 0203676 and DMS - 0402824. J. Sunil Rao is partially supported by NSF grants DMS - 0203724, DMS - 0405072 and NIH grant K25-CA89868.

Appendix

1. *Proof of Lemma.* Write $g(\lambda) = \hat{Q}_{M,\lambda}$. It can be shown (detail omitted) that $g'(\lambda) = 2\lambda y' B_\lambda A_\lambda B_\lambda' y$, where $A_\lambda = B'(W'W + \lambda BB')^{-1}B$, $B_\lambda = W(W'W + \lambda BB')^{-1}B$ with $B' = (0\ I_q)$ and $W = (X\ Z)$. Hence $g'(\lambda) \geq 0$ for $\lambda > 0$. Also $Q_{M,\lambda} \to \hat{Q}_M$ as $\lambda \to 0$.

2. *Proof of Theorem.* Consider the fence inequality

$$\hat{Q}_{M,\lambda} - \hat{Q}_{\tilde{M},\tilde{\lambda}} \leq c_n, \qquad \text{(A.1)}$$

where $(\bar{M}, \bar{\lambda})$ minimizes $\hat{Q}_{M,\lambda}$. Also consider the fence inequality using $\hat{Q}_M = y' P_{W^\perp} y$, which is

$$\hat{Q}_M - \hat{Q}_{\tilde{M}} \leq c_n. \qquad \text{(A.2)}$$

By Lemma, we must have $\bar{\lambda} = 0$, and $\bar{M} = \tilde{M}$, hence $\hat{Q}_{\bar{M},\bar{\lambda}} = \hat{Q}_{\tilde{M}}$. It follows, again by Lemma, that for the same c_n, (A.2) holds if and only if (A.1) holds for some λ. Therefore, the models within the fence, in terms of p and q, are the same under both procedures. It is then easy to see, according to the selection criterion, that the same model $M_0 = M_0(c_n)$, in terms of p and q, will be selected under both procedures for the given c_n. It then follows that the c_n^* selected using the adaptive procedure will be the same under both procedures. Then, once again using the above argument, the optimal model M_0^*, in terms of p and q, will be the same under both procedures.

The formulae below (7) can be derived using the expressions of BLUE and BLUP (*e.g.*, Jiang 2007, §2.3.1) and the following identity (*e.g.*, Sen and Srivastava 1990, page 275): If U is $n \times q$ and V is $q \times n$, then $(P + UV)^{-1} = P^{-1} - P^{-1}U(I_q + VP^{-1}U)^{-1}VP^{-1}$ so long as the inverses exist.

References

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 80, 28-36.

Chatterjee, S., Lahiri, P. and Li, H. (2007). Parametric bootstrap approximation to the distribution of EBLUP, and related prediction intervals in linear mixed models. *Annals of Statistics*, to appear.

Datta, G.S., and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 10, 613-627.

Datta, G.S., and Lahiri, P. (2001). Discussions on a paper by Efron & Gous. (Ed., P. Lahiri) *Model Selection*, IMS Lecture Notes/Monograph 38.

Fabrizi, E., and Lahiri, P. (2004). A new approximation to the Bayes information criterion in finite population sampling. Technical Report, Dept. of Math., Univ. of Maryland.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Ganesh, N. (2009). Simultaneous credible intervals for small area estimation problems. *Journal of Multivariate Analysis*, in press.

Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science*, 9, 55-93.

Hastie, T., and Tibshirani, R.J. (1990). *Generalized Additive Models*. New York: Chapman and Hall.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. New York: Springer.

Jiang, J., Rao, J.S., Gu, Z. and Nguyen, T. (2008). Fence methods for mixed model selection. *Annals of Statistics*, 36, 1669-1692.

Jiang, J., Nguyen, T. and Rao, J.S. (2009). A simplified adaptive fence procedure. *Statistics and Probability Letters*, 79, 625-629.

Kauermann, G. (2005). A note on smoothing parameter selection for penalized spline smoothing. *Journal of Statistical Planning and Inference*, 127, 53-69.

Laird, N.M., and Ware, J.M. (1982). Random effects models for longitudinal data. *Biometrics*, 38, 963-974.

Meza, J., and Lahiri, P. (2005). A note on the C_p statistic under the nested error regression model. *Survey Methodology*, 31, 105-109.

Morris, C.N., and Christiansen, C.L. (1995). Hierarchical models for ranking and for identifying extremes with applications. *Bayes Statistics* 5, Oxford Univ. Press.

Opsomer, J.D., Breidt, F.J., Claeskens, G., Kauermann, G. and Ranalli, M.G. (2007). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society B*, to appear.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Ruppert, R., Wand, M. and Carroll, R. (2003). *Semiparametric Regression*. Cambridge Univ. Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

Sen, A., and Srivastava, M. (1990). *Regression Analysis*. New York: Springer.

Wand, M. (2003). Smoothing and mixed models. *Computational Statistics*, 18, 223-249.

Wang, J.-L. (2005). Nonparametric regression analysis of longitudinal data. *Encyclopedia of Biostatistics*, 2nd Ed.

Survey Methodology, June 2010
Vol. 36, No. 1, pp. 13-22
Statistics Canada, Catalogue No. 12-001-X

Gross flow estimation in dual frame surveys

Yan Lu and Sharon Lohr [1]

Abstract

Gross flows are often used to study transitions in employment status or other categorical variables among individuals in a population. Dual frame longitudinal surveys, in which independent samples are selected from two frames to decrease survey costs or improve coverage, can present challenges for efficient and consistent estimation of gross flows because of complex designs and missing data in either or both samples. We propose estimators of gross flows in dual frame surveys and examine their asymptotic properties. We then estimate transitions in employment status using data from the Current Population Survey and the Survey of Income and Program Participation.

Key Words: Complex surveys; Dual frame surveys; Jackknife; Longitudinal estimation; Missing data.

1. Introduction

Many current surveys follow the same individuals at regular time intervals so that longitudinal quantities such as transitions in employment status and poverty status can be studied. The U.S. Current Population Survey (CPS; United States Census Bureau 2006), for example, uses a rotating panel design in which persons in a housing unit selected for the survey are interviewed for four consecutive months, rested for eight months, and then interviewed again for four consecutive months. This design allows estimation of quantities related to individuals' changes over time. Since many survey responses are categorical, gross flows, which are transitions among states of a categorical variable over time, are particularly important.

Table 1 displays the counts of a categorical variable measured at two times in a population of N units. At time 1, the variable can be in one of r states and at time 2, the variable can be in one of c states. To illustrate Table 1, we give the following example. In studying changes in employment status, we might have $r = 2$ and $c = 2$, with state 0 representing unemployment and state 1 representing employment. Then X_{00} gives the count of persons in the population who are unemployed at both times, X_{10} is the number of persons who are employed at time 1 but unemployed at time 2, X_{0+} is the total number of persons who are unemployed at time 1, and so on. It is of interest to obtain estimates and standard errors of the gross flows X_{kl}, $k = 0, ..., r-1$, $l = 0, ..., c-1$, using survey data. This can be complicated in practice because of missing data and other problems.

While successive cross-sectional estimates can assess a change in unemployment rates over time, only a longitudinal survey addresses issues such as persistence of unemployment in individuals. Gross flow estimation using survey data has been studied by many authors, including Chambers, Woyzbun and Pillig (1988), Hocking and Oxspring (1971), Blumenthal (1968), Chen and Fienberg (1974), Stasny (1984, 1987), and Stasny and Fienberg (1986). Most of this work considered methods for obtaining maximum likelihood (ML) estimators for expected cell values in contingency tables with partially cross-classified data. Pfeffermann, Skinner and Humphreys (1998) proposed estimators that account for misclassification in survey data. All of this work has assumed that a probability sample, usually a simple random sample, has been taken from a single sampling frame.

Table 1
Gross flow table for population

		Time 2					
		0	**1**	**2**	⋯	$c-1$	
Time 1	**0**	X_{00}	X_{01}	X_{02}	⋯	$X_{0,\,c-1}$	X_{0+}
	1	X_{10}	X_{11}	X_{12}	⋯	$X_{1,\,c-1}$	X_{1+}
	2	X_{20}	X_{21}	X_{22}	⋯	$X_{2,\,c-1}$	X_{2+}
	⋮	⋮	⋮	⋮	⋱	⋮	⋮
	$r-1$	$X_{r-1,\,0}$	$X_{r-1,\,1}$	$X_{r-1,\,2}$	⋯	$X_{r-1,\,c-1}$	$X_{r-1,\,+}$
		X_{+0}	X_{+1}	X_{+2}	⋯	$X_{+,\,c-1}$	N

A number of longitudinal surveys, such as the Canadian National Longitudinal Survey of Children and Youth and the Canadian Household Panel Survey, have now started or are considering implementation of a dual frame or multiple frame design. In a multiple frame survey, probability samples are selected independently from two or more frames. Using more than one frame often gives better coverage of the population, and can achieve considerable cost savings in some populations. For example, the Assets and Health Dynamics Survey (Heeringa 1995), with the goal of estimating characteristics of the population aged over 65, used a dual frame survey in which frame A was

1. Yan Lu, Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM, 87131-0001. E-mail: yljenlu@gmail.com; Sharon Lohr, School of Mathematical and Statistical Sciences, Arizona State University, Tempe AZ 85287-1804. E-mail: sharon.lohr@asu.edu.

the frame for a national general population survey and frame B was a list of Medicare enrollees. The structure of this survey is illustrated in Figure 1. Frame A covered the entire population but required extensive screening to identify individuals in the target population and was thus expensive to sample from; frame B was less expensive to sample, but did not include the entire population. Kalton and Anderson (1986) described uses of dual frame surveys to sample rare populations; Blair and Blair (2006) argued that dual frame surveys can take advantage of less expensive sampling modes such as internet sampling when sampling rare populations.

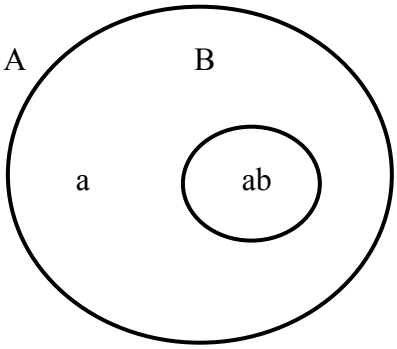

Figure 1 Frame B is a subset of frame A

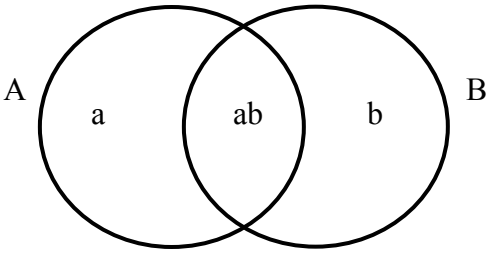

Figure 2 Frames A and B are both incomplete but overlapping

In other situations, both frames may be incomplete, as depicted in Figure 2. Hartley (1962, 1974) first proposed estimators for the dual frame survey design in Figure 2, when independent samples are taken from each frame. Subsequent developments are given in Bankier (1986), Fuller and Burmeister (1972), Skinner and Rao (1996), and Lohr and Rao (2000). Lohr and Rao (2006) summarized methods for estimating population quantities in cross-sectional multiple frame surveys.

In this paper, we propose estimators for gross flows that can be applied to dual frame surveys in which longitudinal information is collected in one or both samples. Units sampled in one or both surveys are followed over time; in some cases, additional units are sampled at later times to incorporate new population units or compensate for attrition. A longitudinal dual frame survey presents additional challenges to those found in longitudinal single frame surveys or in cross-sectional dual frame surveys. Missing data can occur in the sample from either frame, and units may change frame membership between interviews in the survey. In addition, either sampling design may be complex, with stratification and clustering. In an overlapping dual frame survey such as that depicted in Figure 2, one wishes to use the information in the overlap as efficiently as possible. The problem studied in this article is to use all the information sampled from frame A and frame B to estimate the transition probabilities of the population.

The article is organized as follows. In Section 2, we set up the research problem. In Section 3, we derive gross flow estimators in dual frame surveys for complex samples with possibly missing data. In Section 4, we derive asymptotic properties and discuss variance estimation. An application of our research to the Current Population Survey and Survey of Income and Program Participation is given in Section 5. Finally, we give our conclusions in Section 6.

2. Notation and sample quantities

Suppose there are two sampling frames, frame A and frame B, which together cover the population of interest $A \cup B$ as shown in Figure 2. In Hartley's (1962) notation, there are three nonoverlapping domains: $a = A \cap B^c$, $b = A^c \cap B$, and $ab = A \cap B$, where c denotes complement of a set. The population sizes for frames A and B are N_A and N_B, with domain population sizes N_a, N_b, and N_{ab}. We assume that N_A and N_B are known, but the population size $N = N_A + N_B - N_{ab}$ may be unknown. In this article, we assume that both the population and the frames are fixed over time. These are strong assumptions but in many longitudinal surveys the population of interest and the frames may be defined for time 1.

Assume for this section that domain membership is constant over time. For simplicity of notation in this paper we assume that $r = 2$ and $c = 2$ so that there are two possible categories at each time; the general case is similar. Since the three domains are nonoverlapping, each population count X_{kl}, $k = 0, 1$, $l = 0, 1$, can be written as $X_{kl} = X_{kla} + X_{klab} + X_{klb}$, where X_{kld} is the number of population units in domain d that are in state k at time 1 and state l at time 2. The corresponding population and domain probabilities are $p_{kl} = X_{kl}/N$ and $p_{kld} = X_{kld}/N_d$ for $d \in \{a, ab, b\}$.

Independent probability samples, S_A and S_B, with sample sizes n_A and n_B, are taken from frames A and B. Let w_i^A be the weight of sampled unit i for the sample from frame A and let w_j^B be the weight of sampled unit j for the sample from frame B. We may take w_i^A to be the sampling weight $[P(i \in S_A)]^{-1}$ or a Hájek-type weight $[P(i \in S_A)]^{-1} N_A /$ (sum of sampling weights in S_A). Other

weighting schemes for longitudinal data, discussed in Verma, Betti and Ghellini (2007) and Lavallée (2007), might also be used. Let $\mathbf{y}_i = (y_{i1}, y_{i2})$ be the response for unit i in S_A, with $y_{i1}, y_{i2} \in \{0, 1, M\}$ where M denotes that the value is missing. Then $\hat{X}^A_{kla} = \sum_{i \in S_A} w_i^A \, I(y_{i1} = k) \, I(y_{i2} = l) \, I(i \in a)$ and $\hat{X}^A_{klab} = \sum_{i \in S_A} w_i^A \, I(y_{i1} = k) \, I(y_{i2} = l) \, I(i \in ab)$ estimate the population counts for the (k, l) cell in domains a and ab from S_A, for k, $l \in \{0, 1, M\}$. Let $\mathbf{y}_j = (y_{j1}, y_{j2})$ be the response for unit j in S_B, and let $\hat{X}^B_{klb} = \sum_{j \in S_B} w_j^B \, I(y_{j1} = k) \, I(y_{j2} = l) \, I(j \in b)$ and $\hat{X}^B_{klab} = \sum_{j \in S_B} w_j^B I(y_{j1} = k) \, I(y_{j2} = l) \, I(j \in ab)$ be the corresponding estimators from S_B.

In this paper, we assume that domain membership can be determined for every sample unit and that the responses $\mathbf{y}_i$ have no classification error. Thus, we assume that we know whether each unit in the frame A or frame B sample belongs to the other frame or not. We also assume that there is no measurement error for $\mathbf{y}_i$ and $\mathbf{y}_j$ – in the employment example, this means that every respondent gives the correct response for his or her employment status. Thus, the methods we proposed in our article are sensitive to misclassification of observations into domains and into cells. If the domain means differ or if observations are classified incorrectly, the estimators of gross flows could be biased; Pfeffermann *et al.* (1998) discussed methods of accounting for misclassification in single frame surveys.

The estimators from S_A are displayed in Table 2. A similar table may be constructed for the estimators from S_B. We assume that each unit is sampled during one or both time periods. If there is no missing data, then all the estimated counts for cells (k, M) and (M, l) are zero. Using the exact or approximate unbiasedness of the estimators, depending on whether the sampling or Hájek weights are used, when there is no missing data, $E[\hat{X}^A_{kla}] \approx X_{kla}$, $E[\hat{X}^A_{klab}] \approx E[\hat{X}^B_{klab}] \approx X_{klab}$ and $E[\hat{X}^B_{klb}] \approx X_{klb}$.

Table 2
Estimators from the frame A sample

			Time 2			
			0	**1**	**Missing**	
Time 1	**domain *a***	**0**	$\hat{X}^A_{00a}$	$\hat{X}^A_{01a}$	$\hat{X}^A_{0Ma}$	$\hat{X}^A_{0+a}$
		1	$\hat{X}^A_{10a}$	$\hat{X}^A_{11a}$	$\hat{X}^A_{1Ma}$	$\hat{X}^A_{1+a}$
		Missing	$\hat{X}^A_{M0a}$	$\hat{X}^A_{M1a}$		$\hat{X}^A_{M+a}$
	domain *ab*	**0**	$\hat{X}^A_{00ab}$	$\hat{X}^A_{01ab}$	$\hat{X}^A_{0Mab}$	$\hat{X}^A_{0+ab}$
		1	$\hat{X}^A_{10ab}$	$\hat{X}^A_{11ab}$	$\hat{X}^A_{1Mab}$	$\hat{X}^A_{1+ab}$
		Missing	$\hat{X}^A_{M0ab}$	$\hat{X}^A_{M1ab}$		$\hat{X}^A_{M+ab}$
			$\hat{X}^A_{+0}$	$\hat{X}^A_{+1}$	$\hat{X}^A_{+M}$	$\hat{N}_A$

3. Gross flow estimators in dual frame surveys

In this section, we derive gross flow estimators for complex samples in dual frame surveys. A dual frame pseudo-likelihood approach is used to account for the sampling designs and missing data mechanism. A dual frame approach can improve precision of the estimators and provide more flexibility to model the missing data mechanism. Methods in current use for handling missing data are based on standard statistical methods and fall into four general categories (Little and Rubin 2002): complete-case analysis, weighting methods, imputation methods and model-based methods. We adopt a model-based approach for the missing data. In this section, we first consider a simple setup with simple random samples from a population with no missing data. Then we add a model for the missing data mechanism. Finally, we discuss estimators for more complex survey designs.

3.1 Simple random samples with complete data

To motivate the estimator in the general case, we first study estimation of gross flows when there is no missing data and when the sample from each frame is a simple random sample. Then $x^A_{kld} = n_A \hat{X}^A_{kld} / N_A$, for $d = a, ab$, is the observed sample count in cell kl and domain d from S_A; $x^B_{kld} = n_B \hat{X}^B_{kld} / N_B$ for $d = b, ab$ is the corresponding observed sample count from S_B.

If the sampling fractions are small, a multinomial approximation may be used for the likelihood. For the sample from frame A, there are eight cells with associated probabilities $P^A_{kld} = p_{kld} N_d / N_A$, for $k, l \in \{0, 1\}$ and $d \in \{a, ab\}$. The related probabilities for the sample from frame B are $P^B_{kld} = p_{kld} N_d / N_B$ for $k, l \in \{0, 1\}$ and $d \in \{b, ab\}$. Using the multinomial distribution and the assumption that the samples from the two frames are selected independently, the likelihood function is

$$L(\mathbf{p}, N_{ab}) \propto \prod_{k,l,d} (P^A_{kld})^{x^A_{kld}} \times \prod_{k,l,d} (P^B_{kld})^{x^B_{kld}}.$$

Although the likelihood is written for simplicity in terms of P^A_{kld} and P^B_{kld}, the underlying parameters of interest are $\mathbf{p} = (p_{00a}, p_{01a}, ..., p_{11b})$ and N_{ab}.

Setting the partial derivatives of the loglikelihood with respect to the parameters equal to zero, the maximum likelihood estimators are $\hat{p}_{kla} = x_{kla} / n^A_a$, $\hat{p}_{klb} = x_{klb} / n^B_b$ and $\hat{p}_{klab} = (x^A_{klab} + x^B_{klab}) / (n^A_{ab} + n^B_{ab})$, where $n^A_{ab} = \sum_{i \in S_A} I(i \in ab)$, $n^B_{ab} = \sum_{j \in S_B} I(j \in ab)$, $n^A_a = n_A - n^A_{ab}$ and $n^B_b = n_B - n^B_{ab}$. The MLE for N_{ab}, $\hat{N}_{ab}$, is the smaller root of the quadratic equation

$$[n_A + n_B] \hat{N}^2_{ab} - [n_A N_B + n_B N_A + n^A_{ab} N_A + n^B_{ab} N_B] \hat{N}_{ab} + [n^A_{ab} + n^B_{ab}] N_A N_B = 0. \quad (1)$$

Finally, using the above results, we construct the MLEs for X_{kl} and p_{kl}:

$$\hat{X}_{kl} = (N_A - \hat{N}_{ab})\,\hat{p}_{kla} + \hat{N}_{ab}\hat{p}_{klab} + (N_B - \hat{N}_{ab})\hat{p}_{klb}.$$

$$\hat{p}_{kl} = \frac{(N_A - \hat{N}_{ab})\hat{p}_{kla} + \hat{N}_{ab}\hat{p}_{klab} + (N_B - \hat{N}_{ab})\,\hat{p}_{klb}}{N_A + N_B - \hat{N}_{ab}}.$$

These estimators are the same as those obtained by Skinner (1991). However, Skinner used the approximate normal distribution of the response mean $\bar{y}$ in each domain to obtain the MLEs, while our estimators come from a multinomial model. The multinomial model allows us to include partially classified information from units observed at only one time period, as shown in the next section.

3.2 Simple random samples with missing data

In practice, individuals may appear in the sample at only one of the times. This can occur due to sample attrition (when members of the sample drop out during the course of a study) or other causes. In a rotating panel survey such as the CPS, persons rotating out of the survey at time 1 will not be contacted for time 2 and thus their time-2 employment status will be unknown. In other situations, one of the samples may be cross sectional, in which case all observations are measured at exactly one time.

3.2.1 Model for missing data

Blumenthal (1968), Chen and Fienberg (1974), Stasny (1984, 1987) and Stasny and Fienberg (1986) used a two-phase procedure to model the missing data in a single sample. A model is proposed for the complete data, and then the missing data mechanism is modeled. We extend this procedure to our dual frame structures. One advantage of a dual frame survey is that it provides more flexibility for the missing data models.

First, we assume that if all units were measured at both times, the model in Section 3.1 could be used. For the nonresponse mechanism, assume that each observation in cell (k, l) and domain d from S_A has probability ϕ^A_{kld} of being missing at time 1 and probability ψ^A_{kld} of being missing at time 2. We assume the unit cannot be missing at both times.

This formulation assumes a constant probability that an observation will be missing within a given cell, domain, and frame. If data could be missing for different reasons, additional parameters could be used to distinguish observations that have partial classification because of, say, the rotating panel design, and observations that have partial classification because of nonresponse. In Section 5, we discuss an alternative approach that might be used with multiple mechanisms for missing data.

For $k, l \in \{0, 1\}$, the probability that a unit from S_A is observed in cell (k, l) and domain d is

$$Q^A_{kld} = P^A_{kld}(1 - \phi^A_{kld} - \psi^A_{kld}).$$

The probability that a unit from S_A is observed in cell (k, M) and domain d is

$$Q^A_{kMd} = \sum_{l=0}^{1} P^A_{kld}\psi^A_{kld}.$$

Similarly, the probability that a unit from S_A is observed in cell (M, l) and domain d is

$$Q^A_{Mld} = \sum_{k=0}^{1} P^A_{kld}\phi^A_{kld}.$$

The probabilities for frame B are defined similarly with $Q^B_{kld} = P^B_{kld}(1-\phi^B_{kld}-\psi^B_{kld})$, $Q^B_{kMd} = \sum_{l=0}^{1} P^B_{kld}\psi^B_{kld}$ and $Q^B_{Mld} = \sum_{k=0}^{1} P^B_{kld}\,\phi^B_{kld}$.

Under this two phase model, and using the assumption of independence of the samples, the likelihood function for the two samples is:

$$\begin{aligned} L(\mathbf{p}, \boldsymbol{\psi}, \boldsymbol{\phi}, N_{ab}) \propto & \prod_{k\in\{0,1\}}\prod_{l\in\{0,1\}}\prod_{d\in\{a,\,ab\}} (Q^A_{kld})^{x^A_{kld}} \\ & \times \prod_{k\in\{0,1\}}\prod_{l\in\{0,1\}}\prod_{d\in\{b,\,ab\}} (Q^B_{kld})^{x^B_{kld}} \\ & \times \prod_{k\in\{0,1\}}\prod_{d\in\{a,\,ab\}} (Q^A_{kMd})^{x^A_{kMd}} \\ & \times \prod_{l\in\{0,1\}}\prod_{d\in\{a,\,ab\}} (Q^A_{Mld})^{x^A_{Mld}} \\ & \times \prod_{k\in\{0,1\}}\prod_{d\in\{b,\,ab\}} (Q^B_{kMd})^{x^B_{kMd}} \\ & \times \prod_{l\in\{0,1\}}\prod_{d\in\{b,\,ab\}} (Q^B_{Mld})^{x^B_{Mld}}, \end{aligned} \tag{2}$$

where $\boldsymbol{\psi}$ is the vector of ψ^A_{kld}'s and ψ^B_{kld}'s and $\boldsymbol{\phi}$ is the vector of ϕ^A_{kld}'s and ϕ^B_{kld}'s.

The expression in (2) is for the most general model, in which both surveys are longitudinal and both have missing data at each time period. If frame A uses a rotating panel survey, for example, then all of the probabilities Q^A_{kld} are nonzero: the units in the panels measured at both time periods will be included in the estimators x^A_{kld} for $k, l \in \{0, 1\}$, the units in the panels leaving the survey after time 1 will be included in the estimators x^A_{kMd}, and the units in the incoming panels will be included in the estimators x^A_{Mld}. Depending on the structure of the surveys, some of the factors in (2) may be omitted. For example, if the survey from frame B is a repeated cross-sectional survey with small sampling fraction, the probabilities Q^B_{kld} for $k, l \in \{0, 1\}$ will be close to zero, and we would omit those factors from the likelihood.

The likelihood in (2) can be written as a product of a factor with N_{ab} and a factor containing the remaining parameters. As a consequence, the MLE for N_{ab} is again the smaller root of the equation in (1). We discuss the estimators of the remaining parameters in the next section.

3.2.2 Model identifiability and reduced models

A problem with maximizing the likelihood in (2) is that under the general model there are a total of 42 parameters while the two samples have only 32 observed cell counts. Thus we cannot estimate all the parameters under the most general model. But we can consider models with reduced parameterizations, as done in Chen and Fienberg (1974) for single frame surveys. The dual frame situation, in fact, gives much more flexibility for modeling the missing data because of the independent information from the two samples about domain ab.

We first state conditions for a reduced model to be locally identifiable. Let $\boldsymbol{\theta}$ denote the s-vector of parameters of interest; in our case, $\boldsymbol{\theta}$ would include linearly independent components of $\mathbf{p}$, N_{ab}/N, and parameters for the missing data mechanism. In the likelihood in (2), the probabilities from the independent multinomial samples are Q^A_{kld} and Q^B_{kld}. These probabilities may be written as functions of $\boldsymbol{\theta}$, with $\mathbf{Q}^A(\boldsymbol{\theta}) = (Q^A_{00a}, ..., Q^A_{lMab})$ a g-vector of the nonzero Q^A_{kld}'s and $\mathbf{Q}^B(\boldsymbol{\theta}) = (Q^B_{00b}, ..., Q^B_{lMab})$ a q-vector of the nonzero Q^B_{kld}'s. When all cells in Table 2 and the analogous table for frame B have nonzero probabilities, $g = q = 16$. Let $\mathbf{D} = (\mathbf{D}'_A, \mathbf{D}'_B)'$ be the derivative matrix of the transformation, with $\mathbf{D}_{A(\alpha\beta)} = \partial \mathbf{Q}^A_\alpha / \partial \theta_\beta$ and $\mathbf{D}_{B(\delta\beta)} = \partial \mathbf{Q}^B_\delta / \partial \theta_\beta$ for $\alpha = 1, ..., g-1$, $\delta = 1, ..., q-1$, and $\beta = 1, ..., s$. Then, using Theorems 3, 4 and 5 in Catchpole and Morgan (1997), the model is locally identifiable if the matrix $\mathbf{D}$ is of full rank. The proof for the dual frame situation is given in Lu (2007).

In a dual frame survey, we consider two types of models for the missing data. In a Type (1) model, the probabilities of missing time-1 or time-2 information for cell (k, l) is the same for each domain within a frame, *i.e.*, $\phi^A_{kla} = \phi^A_{klab} = \phi_{klA}$, $\psi^A_{kla} = \psi^A_{klab} = \psi_{klA}$, $\phi^B_{klb} = \phi^B_{klab} = \phi_{klB}$ and $\psi^B_{klb} = \psi^B_{klab} = \psi_{klB}$. In this type of model, we estimate the ϕ's and ψ's separately from each sample. It might be considered when the samples from the two frames are collected using different modes. For example, if the frame A sample is a mail survey and the frame B sample is a cell phone survey, one might expect different probabilities of dropout from the two samples.

In a Type (2) model, the probabilities of having missing data are the same in each domain, *i.e.*, $\phi^A_{klab} = \phi^B_{klab} = \phi_{klab}$. This type of model might be considered when nonresponse is expected to be related to the cell membership, and frame membership is thought to have little effect on nonresponse. For example, if the two surveys have similar types of designs and administrative procedures, a Type (2) model might be appropriate.

For each type of model, we may need to place additional restrictions on the parameters in order to solve the likelihood equations. Following Stasny and Fienberg (1986) the following are possible restrictions:

$$\begin{aligned}
&\text{Model 1: } \phi_{kl} = \lambda_{t-1(l)},\ \psi_{kl} = \lambda_{t(k)} \qquad (3)\\
&\text{Model 2: } \phi_{kl} = \lambda_{t-1},\ \psi_{kl} = \lambda_t\\
&\text{Model 3: } \phi_{kl} = \lambda_l,\ \psi_{kl} = \lambda_k\\
&\text{Model 4: } \phi_{kl} = \lambda_{t-1(l)},\ \psi_{kl} = \lambda_t\\
&\text{Model 5: } \phi_{kl} = \lambda_{t-1},\ \psi_{kl} = \lambda_{t(k)}.
\end{aligned}$$

Under model 1, the probability that an individual is a nonrespondent in a given time period depends on the given time period and the individual's classification in the observed time period. Under model 2, the probability that an individual is a nonrespondent in a given time period depends only on the given time period. Under model 3, the probability that an individual is a nonrespondent in a given time period depends only on the individual's classification in the observed time period. Under model 4, the probability that an individual is a nonrespondent at time 1 depends on that time period and the individual's classification in the observed month, and the probability that an individual is a nonrespondent at time 2 depends only on the time period 2. Under model 5, the probability that an individual is a nonrespondent at time 1 depends only on the time period, and the probability that an individual is a nonrespondent at time 2 depends on the time period and the individual's classification in the observed month. Many other models are possible in addition to these five models for each type. Using the derivative matrices, it is easily shown that Models 1-5 are all identifiable.

In general, we will not have closed form solutions for the parameter estimates and the parameters must be estimated using an iterative method. We use the function 'nlm' in R (www.r-project.org) to calculate parameter estimates; the code is available from the authors.

3.3 Estimators from complex samples

When either or both samples are collected with a complex design, using the cell counts directly in the likelihood in (2) will give estimators that are not design-consistent. Skinner and Rao (1996) used a pseudo-maximum likelihood (PML) method to obtain design-consistent estimators in cross-sectional dual frame surveys. They showed that, unlike the estimators of Hartley (1962) and Fuller and

Burmeister (1972), the PML estimators for different response variables used the same set of modified weights and thus were internally consistent.

We propose to study estimators inspired by the PML method for gross flows in dual frame longitudinal complex surveys that allow for missing data at either time period in either sample. The basic idea is to use a working assumption of a multinomial distribution from a finite population to give the form of the estimators and use a design effect to adjust the cell counts to reflect the complex survey design.

In the simple random sampling case, x^A_{kld}/n_A is a design-consistent estimator of Q^A_{kld}. To obtain a pseudo-likelihood for general sampling designs, we replace x^A_{kld}/n_A by $\hat{X}^A_{kld}/N_A$, a design-consistent estimator of Q^A_{kld} under the complex sampling design, in the likelihood (2). Define $\bar{x}^A_{kld} = \bar{n}_A \hat{X}^A_{kld}/N_A$ and $\bar{x}^B_{kld} = \bar{n}_A \hat{X}^B_{kld}/N_B$, where, following Skinner and Rao (1996), we allow $\bar{n}_A$ and $\bar{n}_B$ to be arbitrary constants. Note that if N_A or N_B is unknown, it may be estimated by $\hat{N}_A$ or $\hat{N}_B$ instead.

The pseudo-likelihood has the same form as (2), with x^A_{kld}, x^B_{kld}, n_A and n_B replaced by $\bar{x}^A_{kld}$, $\bar{x}^B_{kld}$, $\bar{n}_A$ and $\bar{n}_B$, respectively. Iterative procedures are then used to find the pseudo-MLEs of the quantities of interest p_{kld}, $\boldsymbol{\phi}$, $\boldsymbol{\psi}$ and N_{ab}. By the fact that the pseudo-likelihood factors, $\hat{N}_{ab}$ is found to be the smaller of the roots of

$$\begin{aligned}[\bar{n}_A + \bar{n}_B]\, \hat{N}^2_{ab,\,PML} \\ - [\bar{n}_A N_B + \bar{n}_B N_A + \bar{n}_A \hat{N}^A_{ab} + \bar{n}_B \hat{N}^B_{ab}]\, \hat{N}_{ab,\,PML} \\ + [\bar{n}_A \hat{N}^A_{ab} N_B + \bar{n}_B \hat{N}^B_{ab} N_A] = 0. \end{aligned} \quad (4)$$

In a complex survey, particularly when clustering is involved, the actual sample sizes n_A and n_B do not necessarily reflect the relative amounts of information from the samples. We thus suggest taking $\bar{n}_A$ and $\bar{n}_B$ to be the effective sample size for each sample, with $\bar{n}_A = n_A/$ (design effect of S_A) and $\bar{n}_B = n_B/$ (design effect of S_B). The design effect of an estimator $\hat{\mu}$ is the ratio

$$\frac{[V(\hat{\mu}) \text{ from complex survey design}]}{[V(\hat{\mu}) \text{ from SRS of same size}]}.$$

The design effect is usually different for different variables. For estimating gross flows, however, the only estimators used from the component surveys are estimated cell counts, and we might expect that in many surveys the design effects for the estimators $\hat{X}^A_{kld}$ would all be similar, and would also be similar to the design effect of the estimator $\hat{N}^A_{ab}$. We thus, as in Skinner and Rao (1996), suggest using the design effect for the estimator $\hat{N}^A_{ab}$ in determining $\bar{n}_A$, and the design effect for the estimator $\hat{N}^B_{ab}$ in determining $\bar{n}_B$. If the design effects of the other variables are indeed identical, then the resulting PMLEs will minimize the variances of the estimated quantities; if they differ, the PMLEs will not be optimal but they will be consistent and in most situations will be close to the optimal values (Lohr and Rao 2006). If the design effect for $\hat{N}^A_{ab}$ is unavailable, as would occur, for example, if the survey were poststratified to N^A_{ab}, then we suggest using a generalized design effect, computed by taking an average or weighted average of design effects from other variables in the survey.

4. Properties of the estimators

In this section, we will investigate properties of the estimators. We derive asymptotic variances, discuss jackknife variance estimators, and perform a small simulation study to explore the properties.

4.1 Properties

We consider the general case in which stratified multistage samples are taken from each frame. The estimators of population totals are the standard Horvitz-Thompson or Hájek estimators from complex surveys. From frame A, the parameter vector $\boldsymbol{\eta}_A = [(\mathbf{Q}^A)', N_{ab}/N_A]'$ is estimated by $\hat{\boldsymbol{\eta}}_A = [(\hat{\mathbf{Q}}^A)', \hat{N}^A_{ab}/N_A]'$, where $\hat{Q}^A_{kld} = \hat{X}^A_{kld}/N_A$; similarly, $\boldsymbol{\eta}_B = [(\mathbf{Q}^B)', N_{ab}/N_B]'$ is estimated by $\hat{\boldsymbol{\eta}}_B = [(\hat{\mathbf{Q}}^B)', \hat{N}^B_{ab}/N_B]'$ with $\hat{Q}^B_{kld} = \hat{X}^B_{kld}/N_B$.

Theorem 1: Let $\hat{\boldsymbol{\eta}} = (\hat{\boldsymbol{\eta}}'_A, \hat{\boldsymbol{\eta}}'_B)'$ and $\boldsymbol{\eta} = (\boldsymbol{\eta}'_A, \boldsymbol{\eta}'_B)'$. Assume that the regularity conditions on the inclusion probabilities in Isaki and Fuller (1982) hold for each sample. Let $\tilde{n}_A$ and $\tilde{n}_B$ be the number of primary sampling units in frames A and B, respectively, and let $\tilde{n} = \tilde{n}_A + \tilde{n}_B$. Assume that $\tilde{n}_A$ and $\tilde{n}_B$ both increase such that $\tilde{n}_A/\tilde{n}_B \to \gamma$ for some $0 < \gamma < 1$. Then $\hat{\boldsymbol{\eta}}$ is consistent for $\boldsymbol{\eta}$, and

$$\tilde{n}^{1/2}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \xrightarrow{d} N(0, \boldsymbol{\Sigma}), \quad (5)$$

where $\boldsymbol{\Sigma}$ is a block-diagonal matrix with blocks $\boldsymbol{\Sigma}_A$ and $\boldsymbol{\Sigma}_B$, $\boldsymbol{\Sigma}_A$ is the asymptotic covariance matrix of $\tilde{n}^{1/2}\hat{\boldsymbol{\eta}}_A$ and $\boldsymbol{\Sigma}_B$ is the asymptotic covariance matrix of $\tilde{n}^{1/2}\hat{\boldsymbol{\eta}}_B$. If, in addition, it is assumed that $N_{ab}/N \to \kappa$ for some $0 < \kappa < 1$ and that the model is identifiable, then $\hat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$, the parameter of interest, consists of components of $\mathbf{p}$, N_{ab}/N, $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$, and $\hat{\boldsymbol{\theta}}$ is the pseudo-maximum likelihood estimator of $\boldsymbol{\theta}$. Furthermore, $\tilde{n}^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normal with mean 0 and asymptotic variance $\mathbf{H}_A\boldsymbol{\Sigma}_A\mathbf{H}'_A + \mathbf{H}_B\boldsymbol{\Sigma}_B\mathbf{H}'_B$, where $\mathbf{H}_F$ is the derivative matrix of the function $\boldsymbol{\theta}$ with respect to the parameters $\boldsymbol{\eta}_F$ for frames $F \in \{A, B\}$.

Proof. With gross flows, observed values of all variables are 0 or 1. Thus the boundedness conditions in Lemmas 1 and 2 of Isaki and Fuller (1982) are met, and the estimators of frame A are consistent and asymptotically normal with

$$\tilde{n}^{1/2}_A(\hat{\boldsymbol{\eta}}_A - \boldsymbol{\eta}_A) \xrightarrow{d} N[0, (\gamma/(1+\gamma))\boldsymbol{\Sigma}_A].$$

The same argument applies to give consistency and asymptotic normality for the vector of estimators from frame B, with

$$\tilde{n}_B^{1/2}(\hat{\boldsymbol{\eta}}_B - \boldsymbol{\eta}_B) \xrightarrow{d} N[0,(1-(\gamma/(1+\gamma)))\,\boldsymbol{\Sigma}_B].$$

Combining these two asymptotic results, and using the independence of the sampling designs along with Slutsky's theorem, gives (5). The limiting distribution of $\tilde{n}^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ follows by the delta method, since the parameters in $\boldsymbol{\theta}$ are all twice continuously differentiable functions of those in $\boldsymbol{\eta}$. Since the parameter estimators cannot always be defined explicitly as a function of other statistics from the sample, we may derive the matrices $\mathbf{H}_A$ and $\mathbf{H}_B$ by linearizing the score equations (Binder 1983). The assumption that $N_{ab}/N \to \kappa \in (0,1)$ guarantees that the linearization is well-defined.

Theorem 1 shows that linearization can be used to estimate the variances of parameters of interest. In many situations, however, the matrices $\mathbf{H}_A$ and $\mathbf{H}_B$ are high-dimensional and the linearized variance estimators have complex form. A practical way to estimate the variances of the estimators is to use the jackknife estimator proposed by Lohr and Rao (2000). Under the regularity conditions in their Theorem 4, the jackknife and linearization variance estimators are asymptotically equivalent. The form of the jackknife variance estimator is $v_{JK}(\hat{\boldsymbol{\theta}}) = v_A(\hat{\boldsymbol{\theta}}) + v_B(\hat{\boldsymbol{\theta}})$, where v_A is a jackknife estimator obtained by deleting one primary sampling unit at a time from frame A while using the full data set for frame B, and v_B is a jackknife estimator obtained by deleting one primary sampling unit at a time from frame B while using the full data set for frame A.

4.2 Simulation study

Theorem 1 shows that the dual frame estimators are consistent for the corresponding population quantities under the modeled missing data mechanism. We performed a small simulation study to investigate properties for moderate sample sizes with overlapping frames. We generated the data following the simulation study in Skinner and Rao (1996), with $\gamma_a = N_a/N$ and $\gamma_b = N_b/N$. A cluster sample from frame A was generated with $\tilde{n}_A$ psus and m observations in each psu, and a simple random sample of n_B observations was generated for frame B. We generated the clustered binary responses for the sample from frame A by generating correlated multivariate normal random vectors and then using the probit function to convert the continuous responses to binary responses.

After generating the sample, we calculated the estimators of the probabilities of the union of frame A and frame B, average of the absolute value of the bias and empirical mean squared error (EMSE) under different settings. The EMSE of a given estimator, $\hat{Y}$ is calculated as:

$$\text{EMSE} = \frac{1}{R}\sum_{r=1}^{R}(\hat{Y}_r - Y)^2, \qquad (6)$$

where $\hat{Y}_r$ is the value of $\hat{Y}$ for the r^{th} simulation run. In our simulation study, we used $R = 100$.

The simulation study was performed with factors: (1) γ_a: 0.2 or 0.4, (2) γ_b: 0.2 or 0.4, (3) clustering parameter ρ: 0.3, (4) missing data mechanism: the probability that an individual is a nonrespondent in a given month depends on the time period and the individual's classification in the observed period; or missing completely at random, (5) amount of missing data: close to 10% or close to 20%, (6) sample sizes: $\tilde{n}_A$: 10, 100 or 500; m: 5, n_B: 100, 1,000 or 5,000. All runs used probability parameters $\mathbf{p}_a$: (0.3, 0.1, 0.2, 0.4), $\mathbf{p}_{ab}$: (0.3, 0.1, 0.1, 0.5), and $\mathbf{p}_b$: (0.4, 0.1, 0.1, 0.4). Table 3 shows the results of the simulation study with missing data generated under Model 1 and fitted with both Model 1 and the model using complete records only.

Table 3
Results from the simulation study for missing data generated under Model 1. Case (1) fits the correct model: Model 1; Case (2) uses complete records only. Bias is the average absolute bias for the population gross flow proportions p_{kl}; EMSE is the average empirical mean squared error for the p_{kl}; the proportions used to generate the missing data are $\lambda_{(t-1)0} = 0.141$, $\lambda_{(t-1)1} = 0.070$, $\lambda_{(t)0} = 0.137$ and $\lambda_{(t)1} = 0.068$. Here, $\tilde{n}_A$ is the number of psus in sample A with psu size 5 and n_B is the number of elements in sample B

$\tilde{n}_A$	n_B		p_{00}	p_{01}	p_{10}	p_{11}
10	100	Estimator	0.311	0.120	0.149	0.420
Case 1		Bias	0.040	0.029	0.029	0.040
		EMSE	0.002	0.001	0.001	0.002
			$\lambda_{t-1(0)}$	$\lambda_{t-1(1)}$	$\lambda_{t(0)}$	$\lambda_{(t)}$
		Estimator	0.159	0.095	0.146	0.094
		EMSE	0.001	0.001	0.002	0.001
10	100	Estimator	0.286	0.120	0.146	0.448
Case 2		Bias	0.048	0.029	0.029	0.041
		EMSE	0.004	0.001	0.001	0.002
100	1,000	Estimator	0.321	0.092	0.138	0.449
Case 1		Bias	0.015	0.011	0.009	0.015
		EMSE	3.337e-04	1.798e-04	1.418e-04	3.256e-04
			$\lambda_{t-1(0)}$	$\lambda_{t-1(1)}$	$\lambda_{t(0)}$	$\lambda_{(t)}$
		Estimator	0.145	0.074	0.123	0.068
		EMSE	2.642e-04	9.389e-05	3.917e-04	8.206e-05
100	1,000	Estimator	0.293	0.092	0.135	0.480
Case 2		Bias	0.0280	0.011	0.010	0.040
		EMSE	0.001	1.839e-04	1.711e-04	0.002
500	5,000	Estimator	0.321	0.093	0.135	0.452
Case 1		Bias	0.006	0.008	0.007	0.012
		EMSE	4.960e-05	7.162e-05	6.381e-05	1.857e-04
			$\lambda_{t-1(0)}$	$\lambda_{t-1(1)}$	$\lambda_{t(0)}$	$\lambda_{(t)}$
		Estimator	0.140	0.071	0.123	0.064
		EMSE	4.466e-05	1.818e-05	2.288e-04	3.545e-05
500	5,000	Estimator	0.292	0.092	0.132	0.483
Case 2		Bias	0.028	0.008	0.008	0.043
		EMSE	8.265e-04	7.642e-05	9.571e-05	1.906e-03

When data are missing at random, all models give estimators of the gross flow proportions p_{kl} that are approximately unbiased so we do not report the results here. From Table 3, both the correct model and the analysis of complete records only produce biased estimators of the p_{kl}'s. With larger sample sizes, however, the bias persists in the analysis that uses complete records only, while it diminishes when Model 1 is fit. This example has relatively small probabilities of missing data. With larger amounts of missing data, the contrast between the estimators is more pronounced.

5. Application

In this section, we apply our results to data from the Survey of Income and Program Participation (SIPP) and the Current Population Survey (CPS) within Arizona. Both CPS and SIPP are longitudinal stratified multistage panel surveys. We treat SIPP and CPS as a dual frame survey with the same target population: the Arizona population 18 years old to 64 years old. Using information from both surveys, we want to model the transition probabilities of employment status changes from January 2001 to January 2002 of people between 18 years old and 64 years old. Note that, strictly speaking, these two surveys are not designed as a dual frame survey. They use different questions for the labor force variables. Although we recoded the variables according to the labor force definitions in CPS, it is possible that these different question wordings and orderings produce bias when combining the information. We use this as an example because a real longitudinal dual frame data is not available. Nevertheless, the example shows the potential gains in efficiency by combining the information from two surveys in estimating gross flows.

Both surveys have target population the noninstitutionalized civilian population of the United States. We consider a subset of the population: the population in the labor force from 18 years old to 64 years old. So $N_A = N_B = N_{ab}$ and the estimation problem is a special case of the theory given in Section 3. The longitudinal file for the 2001 and 2002 SIPP (Westat 2001) uses one panel. We merged Wave 1 (where January 2001 records are stored), Wave 4 (where January 2002 records are stored) and the longitudinal weight file, in which the weights are adjusted to sum to the population count. Since the longitudinal panel weights have been adjusted for the nonresponse, we consider this as a no missing data case. The resulted weighted gross flow table from SIPP is given in Table 4.

For the CPS, the rotation group design introduces partially classified data. January 2001 and January 2002 have 50 percent of the sample in common. We use these 50% of the data together with the partially classified data to perform the analysis. The weight variable we use is a cross-sectional weight with cross-sectional nonresponse and calibration adjustments (United States Census Bureau 2006). For individuals present in the survey for only one of the years, we use the weight from that year. For persons present in both Jan 2001 and Jan 2002, we use the average of the two weights. The rule that we chose the average of the two weights is to minimize the variance of the composite estimator. The population group we used is the 18-64 age group, and we excluded persons who were not in that category during both years. The weighted gross flow table from CPS is in Table 5.

Table 4
Gross flow table for SIPP, in Arizona

		Jan 2002		
		Employed	Unemployed	
January 2001	Employed	2,491,029	73,204	
	Unemployed	30,698	30,160	
				2,625,091

Table 5
Gross flow table for CPS, in Arizona

		January 2002			
		Employed	Unemployed	Missing	
January 2001	Employed	1,129,656	38,848	689,497	
	Unemployed	41,586	8,211	36,041	
	Missing	606,549	57,549		
					2,607,937

Since SIPP is considered as a no missing data case, we assumed $\phi_{kl} = \psi_{kl} = 0$ and use a Type 1 model in the data analysis. We adjusted each weight in the CPS data by the factor 2,625,091/2,607,937 to reach a single population total between the two time periods and a single population total between the two surveys. The number of observations in SIPP (frame A) after combining January 2001 and January 2002 are 551 and the design effect for unemployment is about 1.76, so $\overline{n}_A = 551/1.76 = 313$. The design effect for unemployment in CPS (frame B) is about 1.229, so $\overline{n}_B = 1{,}020/1.229 = 830$. Because the likelihood factors, the estimated parameters of probabilities from the five models (3) are all the same. We list the estimated probabilities and the standard errors from SIPP, CPS and data combining these two surveys in Table 6.

Table 6
Estimated transition probabilities using SIPP, CPS, and the dual frame method with SIPP and CPS. Standard errors are given in parentheses

	p_{00}	p_{01}	p_{10}	p_{11}
SIPP	0.9489	0.0279	0.0117	0.0115
	(0.0124)	(0.0093)	(0.0061)	(0.0060)
CPS	0.9088	0.0454	0.0353	0.0106
	(0.0100)	(0.0072)	(0.0064)	(0.0035)
SIPP and CPS	0.9230	0.0381	0.0262	0.0127
	(0.0080)	(0.0058)	(0.0050)	(0.0030)

Due to confidentiality issues, no clustering information is available in the CPS public-use data sets. We used a product of the published design effect and the variance from multinomial sampling to estimate the variances from both SIPP and CPS data. The result from Theorem 1 was applied to estimate the variances of $\hat{p}_{kl}$ for $k, l = 0, 1$. In this special situation, the variance estimate from the combination of the two data sets is reduced to $(\overline{n}_A/(\overline{n}_A + \overline{n}_B))^2 V_A + (\overline{n}_B/(\overline{n}_A + \overline{n}_B))^2 V_B$, where V_A denotes the variance estimate from SIPP data and V_B denotes the variance estimate from CPS data. Table 6 shows that the standard errors are reduced by using the dual frame method.

We also performed goodness-of-fit tests, developed in Lu (2007), for the five models in (3). The parameter estimates from the five models and results from the goodness-of-fit tests, are listed in Table 7. All five models fit the data well, so we recommend adopting the simplest model, Model 3, for the data.

Table 7
Estimated parameters and results of goodness of fit tests

	Estimated Parameters				df	Corrected G^2	p-value
Model 1	$\lambda_{t-1(0)}$	$\lambda_{t-1(1)}$	$\lambda_{t(0)}$	$\lambda_{t(1)}$	3	3.03	0.39
	0.246	0.395	0.277	0.302			
Model 2	λ_{t-1}	λ_t			5	8.58	0.12
	0.255	0.278					
Model 3	λ_0	λ_1			5	6.61	0.25
	0.262	0.353					
Model 4	$\lambda_{t-1(0)}$	$\lambda_{t-1(1)}$	λ_t		4	4.10	0.39
	0.246	0.397	0.278				
Model 5	λ_{t-1}	$\lambda_{t(0)}$	$\lambda_{t(1)}$		4	6.74	0.15
	0.255	0.277	0.313				

With the limited information available on the public-use data sets, we used simple weight adjustments to make the estimated population counts consistent with known totals. The SIPP and CPS weights in the data sets have already been calibrated and adjusted for nonresponse, so that the models for missing data mostly reflect the rotating panel design rather than attrition due to moving and other activities that might be related to employment status.

Future research on these models might include using different weighting adjustments for the longitudinal surveys. In addition, different parameters could be used to distinguish observations that have partial classification because of the rotating panel design, and observations that have partial classification because of nonresponse. To do so, we could introduce a Markov Chain model similar to the one proposed by Stasny (1987). In the complete data model, individuals are allocated to the table according to a single multinomial distribution. At the second step of the process, which is also unobserved, each individual may be chosen to either rotate out of the sample after the interview for month $t-1$ or rotate into the sample before the month t interview according to the sampling plan. Finally, in the third step of the process, each remaining individual may either lose its row classification or lose its column classification by other reasons. Using this model, we can model the nonresponse at both times (*i.e.*, lose both the row and the column classifications).

6. Conclusions

In this article, we developed statistical methods for estimating gross flows from dual frame surveys. These methods are necessary to estimate changes in poverty status or employment status over time. We developed pseudo-maximum likelihood estimators that use the dual frame structure and the properties of the two survey designs. Our models also account for effects of missing data when an individual drops out of the survey or when a rotation panel design is used, so they allow full use of partial information that may be provided by some households. We use a jackknife method to estimate the variance of estimators and examine the properties of the estimators. The results have been applied to real datasets.

In this paper, the categories of the gross flow tables are defined independently from the sample outcomes. It is also possible to define the categories based on values that depend on the sample. For example, in social surveys, the poverty line might be defined using a percentile from the sample and the categories defined as "Below the poverty line" and "Above the poverty line." Methods from this paper can be used to estimate gross flows if the category definitions depend on the sample, but the variance estimators need to account for the effect of estimating the category boundaries.

Although the results in this paper are for dual frame surveys, the methods are general and could be extended to more than two surveys using PML estimators developed in Lohr and Rao (2006). As the number of frames increases, however, so does the complexity of possible missing data mechanisms. Misclassification error may also be more prevalent with a larger number of frames.

Our research is done in the context of survey sampling, but it also applies to other settings in which data could be combined from two independent sources. As it becomes increasingly difficult for a single survey to cover the entire population of interest, we believe these methods for estimating gross flows can provide better coverage of the population with less expense. They also allow for supplementing a general population survey with surveys of specific subpopulations of interest.

Acknowledgements

This research was partially supported by the National Science Foundation under grants SES-0604373 and

DLS-0909630. The authors thank the associate editor and referees for their insightful and helpful comments.

References

Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Blair, E., and Blair, J. (2006). Dual frame web-telephone sampling for rare groups. *Journal of Official Statistics*, 22, 211-220.

Blumenthal, S. (1968). Multinomial sampling with partially categorized data. *Journal of the American Statistical Association*, 63, 542-551.

Catchpole, E.A., and Morgan, B.J.T. (1997). Detecting parameter redundancy. *Biometrika*, 84, 187-196.

Chambers, R.L., Woyzbun, L. and Pillig, R. (1988). Maximum likelihood estimation of gross flows. *Australian Journal of Statistics*, 30, 149-162.

Chen, T., and Fienberg, S.E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, 30, 629-642.

Fuller, W.A., and Burmeister, L.F. (1972). Estimators for samples selected from two overlapping frames. In *Proceedings of the Social Statistics Section*, American Statistical Association, 245-249.

Hartley, H.O. (1962). Multiple frame surveys. In *Proceedings of the Social Statistics Section*, American Statistical Association, 203-206.

Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā*, Series C, 36, 99-118.

Heeringa, S.G. (1995). Technical description of the assets and health dynamics (ahead) survey sample design. Technical Paper, Institute for Social Research, University of Michigan, hrsonline.isr.umich.edu/docs/userg/AHDSAMP.pdf.

Hocking, R.R., and Oxspring, H.H. (1971). Maximum likelihood estimation with incomplete multinomial data. *Journal of the American Statistical Association*, 66, 65-70.

Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression super-population model. *Journal of the American Statistical Association*, 77, 89-96.

Kalton, G., and Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society*, Series A, 149, 65-82.

Lavallée, P. (2007). *Indirect Sampling*. New York: Springer-Verlag.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.

Lohr, S.L., and Rao, J.N.K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.

Lohr, S.L., and Rao, J.N.K. (2006). Estimation in Multiple-frame Surveys. *Journal of the American Statistical Association*, 101, 1019-1030.

Lu, Y. (2007). Longitudinal estimation in dual frame surveys. *Ph.D Dissertation, Arizona State University*.

Pfeffermann, D., Skinner, C. and Humphreys, K. (1998). The estimation of gross flows in the presence of measurement error using auxiliary variables. *Journal of the Royal Statistical Society*, Series A, 161, 13-32.

Skinner, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.

Skinner, C.J., and Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.

Stasny, E.A. (1984). Some stochastic models for estimating gross flows in the presence of nonrandom nonresponse. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, 25-40.

Stasny, E.A. (1987). Some Markov-Chain models for nonresponse in estimating gross labor force flows. *Journal of Official Statistics*, 4, 359-73.

Stasny, E.A., and Fienberg, S.E. (1986). Estimating gross flows using panel data with nonresponse: An example from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 81, 42-47.

United States Census Bureau (2006). Current Population Survey: Design and Methodology. Technical Paper 66, U.S. Census Bureau, Washington, DC.

Verma, V., Betti, G. and Ghellini, G. (2007). Cross-sectional and longitudinal weighting in a rotational household panel: application to EU-SILC. *Statistics in Transition*, 8, 5-50.

Westat (2001). Survey of Income and Program Participation Users' Guide (Supplement to the Technical Documentation). Technical report, Washington, DC.

Survey Methodology, June 2010
Vol. 36, No. 1, pp. 23-34
Statistics Canada, Catalogue No. 12-001-X

Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling

Qixuan Chen, Michael R. Elliott and Roderick J.A. Little [1]

Abstract

We propose a Bayesian Penalized Spline Predictive (BPSP) estimator for a finite population proportion in an unequal probability sampling setting. This new method allows the probabilities of inclusion to be directly incorporated into the estimation of a population proportion, using a probit regression of the binary outcome on the penalized spline of the inclusion probabilities. The posterior predictive distribution of the population proportion is obtained using Gibbs sampling. The advantages of the BPSP estimator over the Hájek (HK), Generalized Regression (GR), and parametric model-based prediction estimators are demonstrated by simulation studies and a real example in tax auditing. Simulation studies show that the BPSP estimator is more efficient, and its 95% credible interval provides better confidence coverage with shorter average width than the HK and GR estimators, especially when the population proportion is close to zero or one or when the sample is small. Compared to linear model-based predictive estimators, the BPSP estimators are robust to model misspecification and influential observations in the sample.

Key Words: Bayesian analysis; Binary data; Penalized spline regression; Probability proportional to size; Survey samples.

1. Introduction

Unequal probability sampling designs are commonly employed in data collection by science and government. Perhaps the simplest unequal probability design is stratified sampling, which samples units from different strata with different inclusion probabilities. Another important form of unequal probability sampling is probability-proportional-to-size (pps) sampling, in which the inclusion probability is proportional to the value of a size variable measured for all population units.

An unequal probability sampling design such as pps sampling is often used for efficient estimation of population means of continuous variables, for which the variance increases with size of unit. However, inferences about discrete variables are often also of interest in a multipurpose survey (*e.g.*, Lehtonen and Veijanen 1998, Lehtonen, Särndal and Veijanen 2005). In this paper, we focus on methods of inference for finite population proportions from unequal probability sampling designs, based on an auxiliary variable measured for all the units in the population. We use pps sampling as a specific design to illustrate and assess our methods.

The inclusion probabilities play important and somewhat different roles in design-based and model-based inference from unequal probability survey samples (Smith 1976, 1994; Kish 1995; Little 2004). In design-based inference, survey variables are fixed, and inference is based on the distribution of the sample inclusion indicators; the standard design-based approaches to estimation such as the Horvitz-Thompson (HT) estimator (1952) and its extensions weight sampled units by the inverse of their inclusion probabilities. These estimators are design consistent (Isaki and Fuller 1982) and provide reliable inferences in large samples without the need for modeling assumptions. However, these estimators are potentially very inefficient, as illustrated in Basu's (1971) famous elephant example. Also, variance estimation is cumbersome because it requires second-order inclusion probabilities. Corresponding confidence intervals are based on asymptotic theory, and may deviate from nominal levels for moderate or small sample sizes.

Model-based inference predicts values of survey variables in the non-sampled units by including the inclusion probabilities as covariates in the prediction model (Little 2004). Model-based prediction estimators are consistent and efficient under the assumed model, but are subject to bias when the underlying model is misspecified. This limitation motivates the development of flexible statistical models that are more robust to model misspecification. For continuous survey data, Zheng and Little (2003) estimated the finite population total using a nonparametric regression on a penalized spline (*p*-spline) of the inclusion probabilities. We propose here Bayesian *P*-Spline Predictive (BPSP) estimators that are suitable for a binary, as opposed to continuous, outcome. We adopt a Bayesian approach to inference for this model, since Bayesian methods often yield better inference for small sample problems, and are conveniently implemented for our proposed model via the Gibbs' sampler. In this approach, auxiliary variables other than the inclusion probability can also be included in the model, but

1. Qixuan Chen is Assistant Professor, Department of Biostatistics, Columbia University, 722 West 168 Street, New York, NY 10032. E-mail: qc2138@columbia.edu; Michael R. Elliott is Associate Professor and Roderick J.A. Little is Professor, Department of Biostatistics, University of Michigan School of Public Health, 1420 Washington Heights, Ann Arbor, MI 48109. E-mail: mrelliot@umich.edu and rlittle@umich.edu.

the inclusion probability is singled out since modeling of this variable is prone to model misspecification.

We compare the performance of BPSP estimators with Hájek (HK, Horvitz-Thompson-type) estimators and with Generalized Regression (GR) estimators for a binary outcome proposed by Lehtonen and Veijanen (1998). The GR approach is a popular model-assisted modification of the design-based estimators that combines predictions from a model with design-weighted model residuals (Montanari 1998), to yield estimates that are approximately design unbiased.

Zheng and Little (2003; 2005) compared HT, *p*-spline prediction, and GR estimates of the total of a continuous survey variable by simulation. They found that *p*-spline model-based estimators had better root mean squared error than the other methods, and with jackknife standard errors providing superior confidence coverage to HT or GR inferences. We conduct similar comparisons for inference about a population proportion for a binary outcome, and show similar advantages for our BPSP estimator over the HK and GR alternatives.

2. Design-based estimator

Suppose that we have a finite population consisting of N identifiable units. Let Y be the binary survey variable of interest and $p = N^{-1}\sum_{i=1}^{N} Y_i$ be the proportion of the population for which $Y = 1$. Let π_i denote the probability of inclusion for unit i, which is assumed to be known for all units in the finite population before a sample is drawn. An unequal probability random sample s with elements $y_1, \ldots y_n$ is then drawn from the finite population according to the inclusion probabilities $\pi_1, \ldots, \pi_N$. The design-based HK estimator in the discussion of Basu (1971) is defined as

$$\hat{p}_{\mathrm{HK}} = \frac{\sum_{i\in s} y_i/\pi_i}{\sum_{i\in s} 1/\pi_i}. \qquad (1)$$

The variance for $\hat{p}_{\mathrm{HK}}$ can be estimated via linearization of the Yates-Grundy estimator (1953) of totals,

$$\hat{V}_{\mathrm{YG}}(\hat{p}_{\mathrm{HK}}) = \left(\sum_{k\in s} 1/\pi_k\right)^{-2} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{y_i - \hat{p}_{\mathrm{HK}}}{\pi_i} - \frac{y_j - \hat{p}_{\mathrm{HK}}}{\pi_j}\right)^2. \qquad (2)$$

The Yates-Grundy variance estimator requires pairwise inclusion probabilities. When the pairwise inclusion probabilities are not available, as in our simulations, the approximate formula proposed by Hartley and Rao (1962),

$$\pi_{ij} \approx \frac{n-1}{n}\pi_i\pi_j + \frac{n-1}{n^2}(\pi_i^2\pi_j + \pi_i\pi_j^2) - \frac{n-1}{n^3}\pi_i\pi_j\sum_{k=1}^{N}\pi_k^2,$$

has frequently been used. An approximate $1-\alpha$ level confidence interval for the population proportion $\hat{p}_{\mathrm{HK}}$ is then obtained based on the normal approximation.

3. Bayesian *P*-Spline Predictive (BPSP) estimator

Royall (1970) argued for the use of models for finite-population descriptive inferences by predicting the unobserved values based on models, since model-based inferences should be more efficient than design-based inferences. To model the relationship between the binary outcome Y and the continuous inclusion probability π, we need to fit a binary regression of Y on π. Parametric binary regressions, such as the linear or quadratic logistic or probit model, may not be adequate in fitting the data. One solution for this problem of inflexibility is to fit a binary regression on a spline of π by adding some knots. However, too many knots may result in the roughness of model fit. One way to overcome this problem is to retain all of the knots but to constrain their influence, by fitting a binary *p*-spline regression model.

Common methods for modeling a binary outcome are logistic and probit regressions, and they generally give similar results. We choose to adopt probit models in our study for computational convenience. The probit regression model for binary outcomes has an underlying truncated normal regression structure on latent continuous data. If the latent continuous data are known, the parameters in binary *p*-spline regression models can be estimated using standard approaches for normal *p*-spline regression models. In a Bayesian context, the posterior distribution of parameters in the probit *p*-spline model can be computed using Gibbs sampling (Albert and Chib 1993; Ruppert, Wand and Carroll 2003, chapter 16). In contrast, the logistic *p*-spline regression model requires a more complicated computation procedure such as the Metropolis-Hastings algorithm. The computational advantage makes the probit link function more desirable than the logit link function in Bayesian binary *p*-spline regression models.

There are various types of *p*-splines. When applying *p*-splines, we need to make choices on the degree and knot locations, and the basis functions used to present the model. We choose to use the truncated polynomial *p*-splines because they are simple and intuitive. More numerically stable estimators can be obtained using *B*-splines via orthogonalizing the truncated power bases (Eilers and Marx

1996). The probit truncated polynomial *p*-spline regression model has a generalized linear mixed model representation,

$$\Phi^{-1}(E(y_i \mid \beta, b, \pi_i)) = \beta_0 + \sum_{k=1}^{p} \beta_k \pi_i^k + \sum_{l=1}^{m} b_l (\pi_i - k_l)_+^p \quad (3)$$

$$b_l \sim N(0, \tau^2)$$

$$l = 1, ..., m;\ i = 1, ..., n,$$

where $\Phi^{-1}(\cdot)$ denote the inverse CDF of a standard normal distribution, and the constants $k_1 < ... < k_m$ are m selected fixed knots. A function such as $(\pi_i - k)_+^p$ is called a truncated polynomial spline basis function with power p, where $(u)_+^p$ is equal to $\{u \times I(u \geq 0)\}^p$ for any real number u. Since the truncated polynomial spline basis function has $p-1$ continuous derivatives, higher values of p lead to smoother spline functions. By specifying a normal distribution for b, the influence of the m knots is constrained in Model (3), which is equivalent to smooth the splines via the penalized likelihood.

The parameters in Model (3) can be estimated using generalized linear mixed model methods. An alternative Bayesian approach that simplifies computation is to assume weak prior and hyperprior distributions and use Gibbs sampling to obtain draws from the posterior distributions of the parameters as follow: the probit regression model for binary responses has an underlying normal regression structure on latent continuous data; if the latent data are known, the posterior distribution of the parameters can be computed using standard results for normal regression models; and given the posterior distribution of the parameters, the latent continuous data can be simulated from a suitable truncated normal distribution. (Ruppert *et al.* 2003, page 290) The detailed algorithm of Gibbs sampling is in the Appendix. In addition, the Bayesian inference for *p*-spline regression can also been implemented using WinBUGS, the standard Bayesian analysis software (Crainiceanu, Ruppert and Wand 2005).

The posterior distribution of the population proportion is simulated by generating a large number D of draws and using the predictive estimator form $\hat{p}_{\mathrm{PR}}^{(d)} = N^{-1}(\sum_{i \in s} y_i + \sum_{j \notin s} \hat{y}_j^{(d)})$, where $\hat{y}_j^{(d)}$ is a draw from the posterior predictive distribution of the j^{th} non-sampled unit of the binary outcome. The average of these draws simulates the Bayesian *P*-Spline Predictive (BPSP) estimator of the finite population proportion, and is denoted as $\hat{p}_{\mathrm{BPSP}}$, where

$$\hat{p}_{\mathrm{BPSP}} = D^{-1} \sum_{d=1}^{D} \hat{p}_{\mathrm{PR}}^{(d)}. \quad (4)$$

The Bayesian analog of a $100 \times (1-\alpha)\%$ confidence interval for the population proportion is a $100 \times (1-\alpha)\%$ credible interval, which can be formed in a number of different ways. We split the tail area α equally between the upper and lower endpoints in the simulations.

Firth and Bennett (1998) showed that any parametric logistic regression model containing an intercept term and the inverse of inclusion probabilities as a covariate, fitted by ordinary, unweighted maximum likelihood, was "internally bias calibrated" (IBC) for population proportions, and thus yields design consistency. This property is also true for logistic truncated polynomial *p*-spline regression models on the inverse of inclusion probabilities, fitted via penalized likelihood. With the probit link function used instead of the logit link function and fitted via Markov chain Monte Carlo algorithm instead of maximum penalized likelihood, the BPSP estimator may no longer have the IBC property. However, the similarity between the probit model and the logistic model implies that the predictive estimator based on a probit *p*-spline regression model is approximately design-consistent. We believe that obtaining efficient estimates with close to nominal confidence coverage in finite samples is more important than exact design consistency.

4. Generalized Regression (GR) estimator

For the estimation of class frequencies of a discrete response variable, Lehtonen and Veijanen (1998) proposed a GR estimator $\hat{t}_{\mathrm{GR}}$ of the total, which combines the predicted values $\hat{y}_i = \hat{\Pr}(Y_i = 1 \mid \pi_i)$ based on a suitable model and the HT estimator for the residuals $r_i = y_i - \hat{y}_i$ of the sampled units,

$$\hat{t}_{\mathrm{GR}} = \sum_{i=1}^{N} \hat{y}_i + \sum_{i \in s} r_i / \pi_i. \quad (5)$$

The GR estimator in Equation (5) is then used in constructing an estimator for population proportions by dividing by the known population size N (Duchesne 2003),

$$\hat{p}_{\mathrm{GR\_1}} = \frac{1}{N}\left(\sum_{i=1}^{N} \hat{y}_i + \sum_{i \in s} r_i / \pi_i \right). \quad (6)$$

We also consider here another version of the GR estimator for the estimation of finite population proportions, in which the denominator of the bias calibration term for the residuals r_i is the estimated population size $\sum_{i \in s} 1/\pi_i$,

$$\hat{p}_{\mathrm{GR\_2}} = \frac{1}{N} \sum_{i=1}^{N} \hat{y}_i + \left(\sum_{i \in s} r_i / \pi_i \right) \left(\sum_{i \in s} 1/\pi_i \right)^{-1}. \quad (7)$$

For the variance estimate of (6), we use the variance estimator of the estimated total of a discrete response variable, given by Lehtonen and Veijanen (1998), divided by N^2. For the variance estimate of (7), we apply the

Taylor linearization technique (Särndal, Swensson and Wretman 1992, page 182). These two variance estimators are shown in equations (8) and (9), respectively.

$$\hat{V}(\hat{p}_{\mathrm{GR\_1}}) = \frac{1}{N^2}\sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}\frac{r_k}{\pi_k}\frac{r_l}{\pi_l}, \tag{8}$$

$$\hat{V}(\hat{p}_{\mathrm{GR\_2}}) = \left(\sum_{i\in s}1/\pi_i\right)^{-2}\sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}\frac{e_k}{\pi_k}\frac{e_l}{\pi_l}, \tag{9}$$

where $e_k = r_k - (\sum_{i\in s} r_i/\pi_i)(\sum_{i\in s} 1/\pi_i)^{-1}$. These variance estimators also require pairwise inclusion probabilities, which can be approximated by the method of Hartley and Rao (1962).

However, the Hartley and Rao approximation may lead to bias in the variance estimator. Thus, we also consider the jackknife method for variance estimation (Shao and Wu 1989). The sample is stratified into n/G strata each of size G with similar values of inclusion probabilities, and the G subgroups are then constructed by selecting one element at a time from each stratum without replacement (Zheng and Little 2005). Let $\hat{p}_{(g)}$ be the same GR estimators in (6) and (7) calculated from the reduced sample without the elements in the g^{th} subgroup, and let $\bar{p}$ be the average of the G estimators based on the G reduced samples. The jackknife variance estimator of $\hat{p}_{\mathrm{GR}}$ is

$$\hat{V}_{\mathrm{jackknife}}(\hat{p}_{\mathrm{GR}}) = \frac{G-1}{G}\sum_{g=1}^{G}(\hat{p}_{(g)}-\bar{p})^2. \tag{10}$$

A design-weighted logistic regression model on other covariates was used as the assisting model to predict $\hat{y}_i$ in the GR estimators for binary outcomes (Lehtonen and Veijanen 1998; Lehtonen *et al.* 2005). Since our interest here is in comparisons of GR estimators with the BPSP estimator, we apply the estimators (6) and (7) with linear probit regression models and probit *p*-spline models, as described in detail in Section 5. For the GR estimator using a linear probit model as the assisting model, we use the inclusion probability as a covariate as well as a weight in our simulations.

5. Simulation study

5.1 Design of the simulation study

Simulation studies are conducted to study the performance of the BPSP estimator compared with the HK estimator, the GR estimators, and the linear model-based predictive estimators for a variety of populations in pps sampling. We present the simulation results for the following six estimators:

a) HK, the Hájek estimator defined by equation (1).
b) LR, predictive estimator of the form $\hat{p}_{\mathrm{LR}} = N^{-1}(\sum_{i\in s} y_i + \sum_{j\notin s}\hat{y}_j^{\mathrm{LR}})$ with prediction $\hat{y}_j^{\mathrm{LR}}$ obtained with the maximum likelihood predictions from the linear logistic regression model containing a constant term and the reciprocal inclusion probability as the covariate. LR has the IBC property, and hence is design-consistent. LR is exactly the same as its GR estimator in equation (6).
c) PR, predictive estimator of the form $\hat{p}_{\mathrm{PR}} = N^{-1}(\sum_{i\in s} y_i + \sum_{j\notin s}\hat{y}_j^{\mathrm{PR}})$ with prediction $\hat{y}_j^{\mathrm{PR}}$ from the Bayesian linear probit model containing an intercept term and the inclusion probability as the covariate.
d) PR_GR, the GR estimator in equation (7), where $\hat{y}_i$ is the prediction for unit i with unknown parameters replaced by weighted maximum likelihood estimates from the probit model with a constant term and the inclusion probability as the covariate.
e) BPSP, the BPSP estimator defined by equation (4) with $p = 1$ and inverse-gamma prior distribution for τ^2 and using 15 knots.
f) BPSP_GR, the GR estimator in equation (7), where $\hat{y}_i$ is the posterior mean of $\Pr(Y_i = 1 \mid \pi_i)$ from the BPSP model.

We only report the simulation results based on the linear splines for the BPSP estimator, since simulations not shown here suggest that linear splines perform as well as quadratic splines or cubic splines in all the simulation scenarios. We choose two fixed numbers of knots (15 or 30), and place knots at evenly spaced sample percentiles. The choices of knots work well and a number of 15 knots is good enough to catch the curvatures in our simulations. In addition, the GR estimators in (6) perform similarly to the estimators in (7); some differences between these estimators emerge in the real application in Section 6, leading us to prefer (7) over (6).

We simulated two artificial populations of size 2,000, using two different distributions, with sampling rates of 5% and 10%, where the size variable takes the consecutive integer values 71, 72, ..., 2,070. The inclusion probabilities in the population were then calculated as proportional to the size variable, with the maximum value about 30 times the minimum values.

Continuous data Z were first generated from normal distributions with mean structure $f(\pi)$ and constant error variance 0.04. Two different mean structures $f(\pi)$ were simulated: a linearly increasing function (LINUP) $f(\pi_i) = k_1\pi_i$ and an exponential function (EXP) $f(\pi_i) = \exp(-4.64 + k_2\pi_i)$. To make the range of Z similar across different mean structures, k_1 takes values of 3 and 6, and k_2 takes values of 26 and 52, when the sampling rate is

10% and 5%, respectively. Figure 1 plots the two populations. We then generated the binary outcome variable Y_1, where Y_1 is equal to one if Z is less than or equal to its superpopulation 10th percentile, otherwise Y_1 is equal to zero. Similarly, we generated the binary outcomes Y_2 and Y_3 by using the superpopulation 50th and 90th percentiles of Z as cut-off values. The target of inference here is the population proportion with Y equal to one.

In each simulation replicate, a finite population was generated before a sample was drawn, and the true finite population proportion with Y equal to one was calculated and denoted as p. A pps sample was then drawn systematically from a randomly ordered list of the finite population. For each population and sample size combination, 1,000 replicates were obtained and the six estimators were compared in terms of empirical bias, root mean squared error (RMSE), and the non-coverage rate of the 95% confidence /credible interval. Simulation results are presented in Tables 1 through 3. Let $\hat{p}_i$ be an estimate of p_i based on the i^{th} pps sample, the empirical bias and RMSE are defined as follow,

$$\text{Bias} = \frac{1}{1,000}\sum_{i=1}^{1,000}(\hat{p}_i - p_i),$$

$$\text{RMSE} = \sqrt{\frac{1}{1,000}\sum_{i=1}^{1,000}(\hat{p}_i - p_i)^2}.$$

5.2 Simulation results

Figure 2 shows the posterior means of $\Pr(Y_i = 1 \mid \pi_i)$ and 95% credible intervals based on the Bayesian probit linear p-spline model for a random pps sample from the EXP case. The upper left plot is the scatter plot of the continuous variable Z in a pps sample, with three horizontal parallel lines superimposed, representing the superpopulation 10th, 50th, and 90th percentiles, respectively. In the upper right plot, the binary variable Y, defined as 1 if Z is less than or equal to the superpopulation 10th percentile, are plotted with black circles, and the superpopulation $\Pr(Y_i = 1 \mid \pi_i)$ are plotted with a solid black curve. The solid grey curve and two dashed grey curves are the posterior means of $\Pr(Y_i = 1 \mid \pi_i)$ and 95% credible intervals based on the Bayesian probit linear p-spline regression model. The other two plots are similar to the upper right plot, but with superpopulation 50th and 90th percentiles as cut-off values in defining Y. These plots show that the true probabilities of $Y = 1$ fall within the 95% credible intervals, and are close to the posterior means of $\Pr(Y_i = 1 \mid \pi_i)$. We conclude that the Bayesian probit p-spline regression model fits well for the binary outcomes in the nonlinear case.

Table 1 shows the empirical bias ($\times 10^3$) for the six estimators in the two populations generated via LINUP and EXP. Overall the design-based estimators (a, d, and f) are less biased than the model-based estimators (b, c, and e). In the LINUP case, the linear probit regression model is correctly specified, so that the empirical bias of the PR estimators are similar to the empirical bias of the BPSP estimator; while in the EXP case, a nonlinear probit regression is needed to fit the data, and thus the PR estimator is more biased than the BPSP estimator when the true population proportions are 0.1 and 0.5. However, the LR estimator has similar to the BPSP estimator empirical bias because of the IBC property. Compared to the model-based PR and BPSP estimators, the PR_GR and BPSP_GR estimator reduce the bias by adding the bias calibration term. Moreover, no matter which assisting models were used, both GR estimators achieve similar empirical bias.

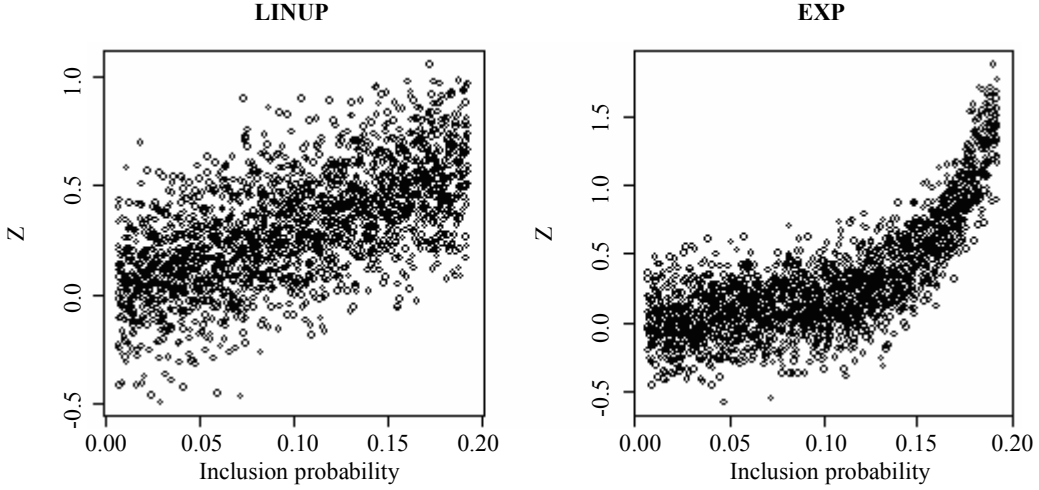

Figure 1 Two simulated artificial populations (N = 2,000)

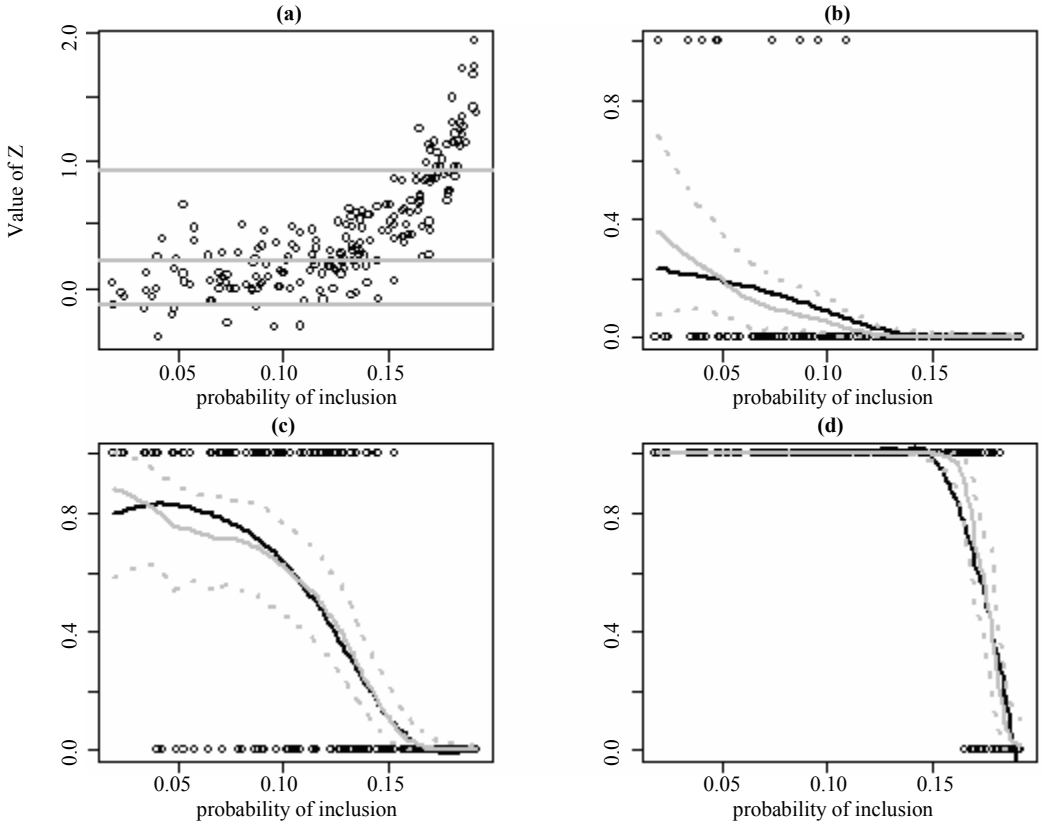

Figure 2 **A random pps sample from the EXP case (n = 200, N = 2,000): (a) scatter plot of *Z*; the three grey lines are the superpopulation 10th, 50th, and 90th percentiles, respectively. (b) black circles are observed units of binary survey variable *Y* in the sample, defined as *Y* = *I* (*Z* ≤ 10th percentile); the grey solid and dashed curves are posterior means of Pr($Y_i = 1|\pi_i$) and 95% credible intervals, respectively, simulated based on a probit *p*-spline model on π; and the black curve is the superpopulation Pr($Y_i = 1|\pi_i$). (c) similar to (b), but with *Y* = *I* (*Z* ≤ 50th percentile). (d) similar to (b), but with *Y* = *I* (*Z* ≤ 90th percentile)**

Table 1
Empirical bias × 1,000 of six estimators (Minimum absolute bias within a row is in italic print)

Population	*n*	True prop.	HK	LR	PR	PR_GR	BPSP	BPSP_GR
LINUP	100	0.10	*-0.01*	13.0	10.3	1.6	8.0	1.2
		0.50	-4.0	*-2.9*	-4.3	-3.0	-5.2	-3.3
		0.90	-0.4	0.3	-2.5	0.3	-2.9	*0.08*
	200	0.10	2.5	7.9	5.8	1.5	5.1	*1.4*
		0.50	3.3	-0.1	-1.3	*-0.06*	-1.7	-0.2
		0.90	1.6	0.4	-1.0	*0.3*	-1.2	*0.3*
EXP	100	0.10	*1.2*	18.1	25.8	4.7	17.0	3.9
		0.50	-4.0	-3.5	12.5	-1.6	*-1.4*	-3.4
		0.90	-1.3	-0.2	-1.0	*-0.1*	-1.0	-0.2
	200	0.10	3.1	11.0	22.1	3.5	13.4	*2.7*
		0.50	3.8	-0.6	14.0	0.4	*0.01*	-0.7
		0.90	2.3	0.1	-0.7	0.1	-0.7	*0.02*

Table 2 shows the empirical root mean squared error ($\times 10^3$) for the six estimators. The BPSP estimator has much smaller empirical root mean squared error than the HK estimator, except when p is 0.1 in the EXP case. Overall the PR estimator performs similarly to the BPSP estimator. To protect again model misspecification, the GR estimators lose some efficiency compared to their corresponding model-based predictive estimators. The PR_GR estimator has similar to the BPSP_GR estimator RMSE, but both of the two GR estimators have smaller RMSE compared to the HK estimator by using assisting models.

Table 3 shows the noncoverage probability ($\times 10^2$) of 95% confidence/credible intervals, the probability that the true finite population proportion is outside the 95% CI of the

estimators. To calculate the variances of estimators, we use the Yates-Grundy variance estimator as defined in equation (2) for the HK estimator; use jackknife resampling method defined by equation (10) for the LR estimator; and use both the linearization (V1) method defined by equation (9) and the jackknife resampling (V2) method for the PR_GR and BPSP_GR estimators. Overall, the confidence coverage of credible interval for the BPSP estimator is closer to the nominal level than the other five estimators, especially when the population proportion p is close to zero or one or when few observations are selected into sample in the tails. Specifically, the BPSP estimator achieves significant improvement in coverage when p is close to zero in both the LINUP and EXP cases, since little data are included in the sample from the lower tail of the two populations. Note that the improved coverage of the BPSP estimator is achieved with intervals that are narrower on average than those of the HK, LR, PR_GR, and BPSP_GR estimators. Similar to the empirical bias and RMSE, the BPSP_GR does not improve the coverage in comparison to the PR_GR estimator by using a flexible assisting model.

The choice of prior and hyperprior distributions in mixed models can have a big effect on inferences. We used a prior distribution $N(0,10^6)$ for the fixed effects parameters, β_i. In our simulations, we report results based on a proper inverse-gamma prior distribution for τ^2, namely $\tau^2 \propto \mathrm{IG}(0.1, 0.1)$. To assess sensitivity to the choice of prior distributions, we also computed results using $\tau^2 \propto \mathrm{IG}(0.01, 0.01)$ and $\tau^2 \propto \mathrm{IG}(0.001, 0.001)$, as well as an improper uniform prior distribution on τ (Gelman 2006). These different priors had little impact on posterior inference of the proportion of interest.

6. Example of tax auditing

We now compare the BPSP estimator with alternative methods on a real population involving income tax auditing data (Compumine 2007). The data set consists of 3,119 Swedish income tax returns for persons who during the year sold mutual funds managed in a foreign country. The outcome of interest Y is whether the income tax return is incorrect (coded as 1 for incorrect, and 0 for correct), and it is measured for all observations in this data set. We treated the 3,119 income tax returns as a finite population here, so that the true population proportion of incorrect income tax returns is 0.517. Since the amount of the realized positive profit is an important feature for determining the amount the tax payer has hidden from taxation for his return of income from the sale of a foreign fund, it was chosen as the size variable used in drawing pps sampling. When the primary measure of interest is the total amount the tax payer has hidden from taxation, it is reasonable to assign a value of 1 Swedish Krona to negative profits, the minimum amount of the positive profits, where negative values are not allowed in the size variable.

One thousand repeated systematic pps samples of size 300 and 600 were drawn without replacement from randomly ordered population lists. The returns with largest profits were included with certainty into the samples of size 300 and 600: there were 78 and 241 such returns respectively. Figure 3 shows that the probability of inclusion has a right-skewed distribution for the population even after excluding the observations with inclusion probability of 1.

We applied the same six estimators as in the simulation study with 30 knots on the pps samples, and compared their performances in terms of empirical bias, RMSE, and average width and noncoverage rate of the 95% confidence/credible interval. For the BPSP estimator, a fixed number of 30 knots are placed at evenly spaced sample percentiles of the inclusion probabilities. For the GR estimators, neither the linearization nor the jackknife variance estimator has predominantly better performance than the other, we present the inference based on the linearization variance estimator for simple calculation. We report the GR estimators based on both equations (6) and (7). The results are displayed in Table 4.

Table 2
Empirical RMSE × 1,000 of six estimators (Minimum RMSE within a row is in italic print)

Population	*n*	True prop.	HK	LR	PR	PR_GR	BPSP	BPSP_GR
LINUP	100	0.10	55.1	57.1	*46.3*	51.3	47.2	51.7
		0.50	65.2	50.8	*47.1*	49.7	47.7	50.0
		0.90	26.3	*22.6*	23.3	22.7	23.5	22.9
	200	0.10	39.3	40.9	*31.8*	36.1	*32.0*	36.2
		0.50	45.7	35.9	*32.8*	34.3	*32.8*	34.6
		0.90	17.8	15.4	15.5	15.4	15.5	*15.3*
EXP	100	0.10	*51.2*	60.1	54.4	51.6	51.8	52.4
		0.50	66.1	56.0	*43.0*	53.2	47.0	51.7
		0.90	24.2	12.4	*12.3*	12.4	*12.3*	*12.3*
	200	0.10	35.9	42.4	39.6	*35.6*	36.0	36.2
		0.50	45.1	38.9	*31.3*	36.1	32.1	35.1
		0.90	15.8	*8.0*	8.1	*8.0*	*8.0*	*8.0*

Table 3
Noncoverage rate of 95% CI × 100 of six estimators (noncoverage rate within a row closest to 5 is in italic print)

Population	*n*	True prop.	HK	LR	PR	PR_GR		BPSP	BPSP_GR	
						V1	V2		V1	V2
LINUP	100	0.10	16.2	18.0	*8.4*	20.9	16.1	9.0	18.4	14.2
		0.50	7.5	9.4	*5.0*	7.2	7.6	4.4	7.3	7.1
		0.90	7.4	11.4	5.7	8.0	9.4	*5.4*	8.4	7.1
	200	0.10	10.8	12.6	6.4	13.9	10.9	*6.2*	12.6	9.4
		0.50	5.5	8.3	5.5	6.2	5.9	*5.1*	6.0	5.5
		0.90	6.0	8.4	4.4	6.1	4.4	*4.7*	6.3	5.5
EXP	100	0.10	15.0	18.1	10.5	19.4	14.8	*9.2*	18.4	14.4
		0.50	*7.4*	13.5	12.2	9.0	11.4	8.9	10.2	8.4
		0.90	*6.1*	10.5	7.9	9.9	7.6	7.0	9.8	7.2
	200	0.10	10.8	13.3	9.9	12.5	11.7	*7.5*	12.4	9.4
		0.50	*6.0*	11.5	14.3	7.2	8.5	6.2	7.5	6.9
		0.90	*5.5*	8.8	5.5	6.8	4.6	*5.5*	6.6	3.7

* V1: variance estimator using linearization; V2: jackknife variance estimator.

Table 4 shows that the BPSP estimator has slightly increased bias but smaller RMSE, shorter average width and closer to the nominal level credible interval than the design-based estimators (a), (d), and (f). Results not shown here indicate that the BPSP estimator with a uniform prior distribution has slightly better performance than that with inverse-gamma prior distribution with respect to empirical bias, RMSE, and coverage rate, because there are more fluctuations in the data and the uniform prior allows the fitted function to have more flexibility. The BPSP_GR estimator is less biased, but achieves less efficiency and worse coverage rate than the BPSP estimator. The predictive estimator using the probit linear regression model as prediction model performs poorly here since the model is misspecified, but its GR estimator does reduce bias and RMSE and improve coverage rate. The BPSP_GR estimator based on equation (6) performs very poorly in terms of RMSE compared to the estimator in equation (7), because a situation similar to that in Basu's (1971) circus elephant example occurs, where one or more observations having very low inclusion probabilities are selected into the sample and hence receive large weights. However, the PR_GR estimator in equation (6) performs as well as that in equation (7) with predictions obtained from the weighted maximum likelihood estimates, where inclusion probability is used as a covariate as well as the sample weights. Overall, the GR estimator in equation (7) is more desirable than that in equation (6). As the sample size increases from 300 to 600, the noncoverage probability of the 95% credible interval of the BPSP estimator approaches the nominal level of 5% quickly from 14% to 5%, but the coverages are consistently below the nominal level for the other estimators.

Compared to the linear model-based predictive estimators, the BPSP estimator is robust not only to model misspecification, but also to the influential observations in the sample. To demonstrate the robustness to the influential observations, we compare the changes in the model fitting using probit *p*-spline models, linear probit model, and quadratic probit model based on the pps sample only in Figure 4, and based on the pps sample as well as the observations with inclusion probabilities of 1 in Figure 5. In each figure, the population is stratified by the 100 quantiles of the probabilities of inclusion, and the true probabilities of $Y = 1$ are calculated and plotted with a black dot for each stratum. The grey curves are the posterior means of $\Pr(Y_i = 1 \mid \pi_i)$ from 10 random pps samples using 3,000-iterate Gibbs sampler and linear spline in the left plot, using linear probit regression in the middle plot, and using quadratic probit regression in the right plot. Figure 4 shows that the probit *p*-spline regression model is more flexible in catching the pattern among the observations than the parametric models. From Figure 4 to Figure 5, the posterior means of $\Pr(Y_i = 1 \mid \pi_i)$ do not change except for those with very large inclusion probabilities using the *p*-spline model. However, the posterior means curves change dramatically using the quadratic probit regression. These comparisons indicate that probit *p*-spline regression model is less likely affected by influential observations, and hence is a good choice of prediction model in the model-based inference.

7. Discussion

Bayesian inferences based on the *p*-spline model outperform the HK estimator, the GR estimators, and linear model-based prediction estimators in our simulations. The BPSP estimators are more efficient than the HK and GR estimators, and despite slightly higher empirical bias, their 95% credible intervals provide better confidence coverage and shorter average interval width, especially when the population proportion is closer to zero or one and few data are selected into the sample in the tails. This suggests the importance of current research in estimating finite population prevalence of rare events.

The BPSP estimator is a natural extension of the regular linear regression model-based estimators of finite population proportions. Compared to linear model-based predictive estimators, the BPSP estimator achieves robustness to model misspecification and influential observations in the sample by using a flexible *p*-spline model, without much loss of efficiency for the sample sizes considered. Therefore, the BPSP estimator is easy to understand while requires complex computation. However, with the availability of WinBUGS, the Bayesian statistical software, the BPSP estimator can be easily implemented by survey practitioners.

Table 4
Comparison of various estimators for empirical bias, root mean squared error, and average width and noncoverage rate of 95% CI, in the tax return example

Methods	bias*100		RMSE*100		average width*100		noncoverage*100	
	300	600	300	600	300	600	300	600
HK	-2.4	-1.8	12.4	10.2	36	29	14.1	10.2
LR	6.7	5.5	11.9	9.2	27	21	43.5	45.6
PR	-11.6	-10.1	12.4	10.6	18	14	69.8	83.4
PR_GR1	-1.2	-0.4	11.5	8.7	31	25	22.4	16.8
PR_GR2	-1.2	-0.3	11.5	8.8	33	26	16.1	11.4
BPSP	-6.8	-2.7	9.3	5.2	27	19	14.2	5.0
BPSP_GR1	-3.0	-0.5	102.6	56.9	77	57	14.4	9.2
BPSP_GR2	-0.7	0.2	12.0	10.1	34	26	15.9	12.8

* GR_1: GR estimators using equation (6);
GR_2: GR estimators using equation (7).

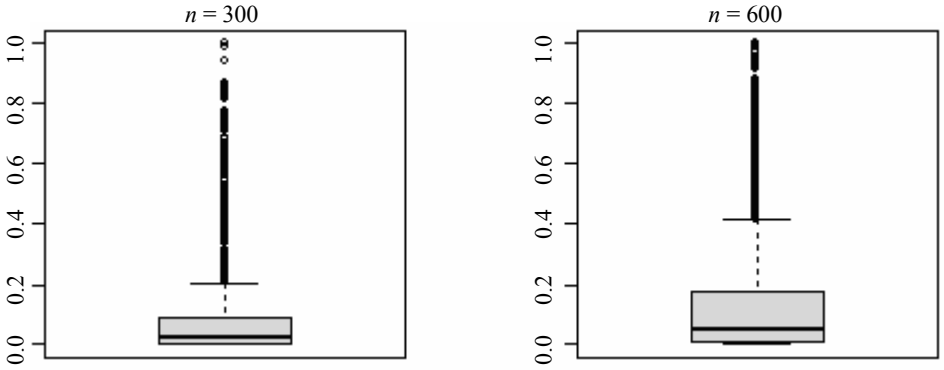

Figure 3 Box plots of the probabilities of inclusion for two sample sizes in the tax auditing example

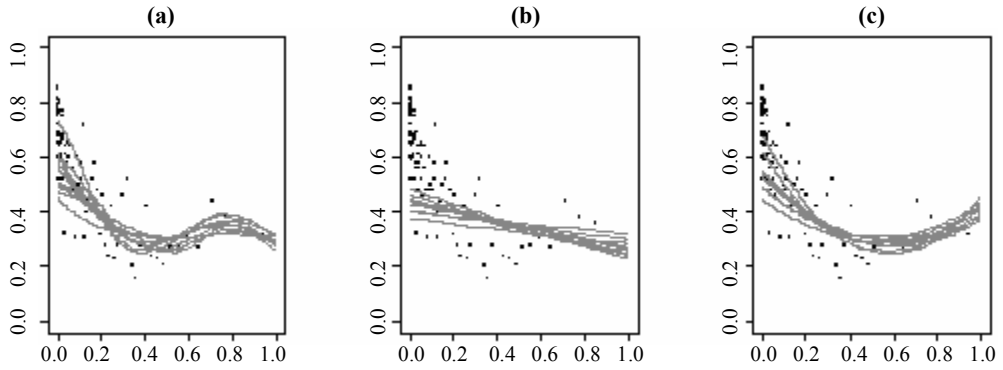

Figure 4 Predictions based on pps samples only in the tax auditing example, *X*-axis: inclusion probabilities π, *Y*-axis: $P(Y = 1|\pi)$; black dots are the true $P(Y = 1|\pi)$ within each percentile of π; grey curves are ten realizations of the posterior means of $P(Y = 1|\pi)$. The prediction models are (a) probit linear *p*-spline regression, (b) linear probit regression, (c) quadratic probit regression

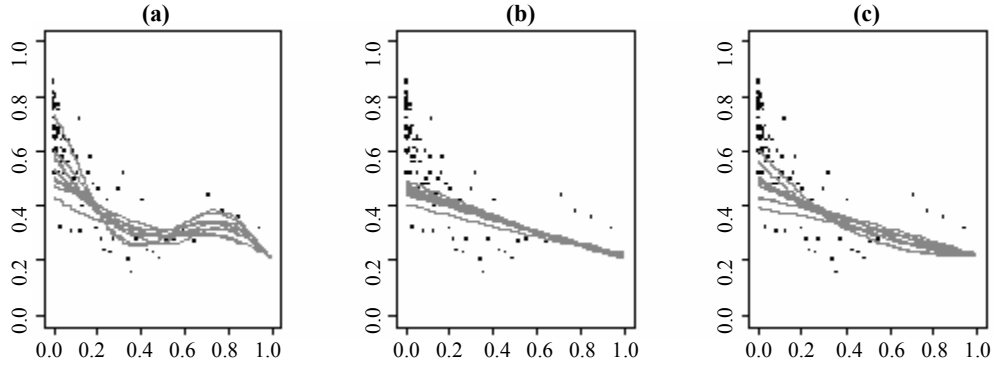

Figure 5 Predictions based on the combined data of pps samples and the observations sampled with certainty in the tax auditing example, *X*-axis: inclusion probabilities π, *Y*-axis: $P(Y = 1|\pi)$; black dots are the true $P(Y = 1|\pi)$ within each percentile of π; grey curves are ten realizations of the posterior mean of $P(Y = 1|\pi)$. The prediction models are (a) probit linear *p*-spline regression, (b) linear probit regression, (c) quadratic probit regression

The BPSP estimators are not sensitive to two choices of prior distributions of τ^2 considered here, though it appears from the tax auditing example that the uniform prior yields slightly smaller bias and RMSE, shorter 95% credible intervals, and better coverage when a nonlinear prediction model is needed. The tax auditing example also shows that in the GR estimator, an estimated population size using the sum of inverse inclusion probabilities is more desirable than the true population size when one or more observations with very low inclusion probability are included in the sample, since the GR estimator with denominator N has high variance and low efficiency in this case.

The design-based estimators and their 95% confidence intervals can provide valid inferences for population proportions when the sample is large. However, these asymptotic properties do not appear to hold when the sample size is moderate or small. The BPSP approach can provide more valid inferences for small samples, especially when the true population proportion to be estimated is close to 0 or 1, although confidence coverage appears to be less than nominal when the sample size gets small, and lack of parsimony of the model is an issue. When estimating proportions away from tails, the BPSP estimator leads to slightly smaller RMSE and closer to the nominal level confidence coverage than the HK and GR estimators, but the improvement is not so significant as in the tails. In this scenario, to avoid the complex computation of the BPSP estimator, the PR_GR estimator based on equation (7) is an alternative to the survey practitioners.

The choice of variance estimator is problematic for some unequal probability designs for the design-based estimators, but the Bayesian *p*-spline prediction approach provides a simulation approximation of the full posterior distribution of the population proportion. Extra work is not needed to estimate the variance or 95% credible interval for the BPSP estimator, as it can be obtained simultaneously with the point estimators. In Zheng and Little (2005), three variance estimators of the *p*-spline model-based estimator for finite population total in a pps sample were compared, including the model-based empirical Bayes variance estimator, the jackknife variance estimate, and the balanced repeated replication (BRR) variance estimate. The simulation studies showed that the jackknife method worked well, whereas the BRR method tended to yield conservative standard errors and the model-based empirical Bayes estimator was vulnerable to misspecification of the variance structure. In the present work, the $1 - \alpha$ level credible interval for the BPSP estimator of population proportion is constructed by splitting α equally between the upper and lower endpoints of the posterior distribution of p. This pure Bayesian approach based on draws from the posterior distributions seems to work well in our setting and avoids the heavy computation associated with the jackknife and BRR method.

The BPSP estimator we propose here can be extended to include additional auxiliary covariates by adding linear terms for these variables. For domain estimation, an interaction term between the spline of inclusion probabilities and the domain indicator should also be modeled. Both the additive effects of auxiliary variables and the interaction between the domain indicator and inclusion probabilities can be represented in a mixed model (Ruppert *et al.* 2003, page 231) and estimated using Gibbs sampling or WinBUGS (Crainiceanu *et al.* 2005). The BPSP estimator for finite population proportions can also be extended to a more general case of a polychotomous response. The Gibbs

sampling approach for the binary case can be generalized to the case of ordered categories, and can be applied to the unordered categories with a latent multinomial distribution (Albert and Chib 1993). Another extension for the BPSP estimator is in the small area estimation, by combing small area random effects with the smooth spline on the inclusion probabilities (Opsomer, Claeskens, Ranalli, Kauermann and Breidt 2008). This extension will be the focus of future research.

Finally, one reviewer questioned whether the proposed approach can be applied in a multipurpose survey with many outcomes, since the modeling procedure does not provide a single set of weights and needs to be repeated for all variables of interest. It is true that our methods are more computationally intensive than existing approaches, but the BPSP method can be easily implemented with a Gibbs sampling algorithm or using WinBUGS, so computing is not a major obstacle. We point out that the simulations in the paper involved repeating the iterative Gibbs analysis 6,000 times, so an equivalent level of computation on a single survey of comparable size would allow the implementation of the BPSP method for 6,000 outcomes! These were done on a garden-variety laptop PC. While we do not advocate automatic use of any analytical method, design or model-based, our point is that computational complexity is no longer a major obstacle to applying these methods. We suggest that the statistical properties of a method are more important than computing time, given modern day computing resources.

Acknowledgements

This work is supported in part by The Dow Chemical Company through an unrestricted grant to the University of Michigan Dioxin Exposure Study. The authors thank the referees and an associate editor for their helpful comments on the original version of this paper.

Appendix

Algorithm of Gibbs sampling

Model (3) can also be written in the matrix form,

$$\Phi^{-1}(E(y_i \mid \beta, b, X, Z)) = (X\beta + Zb)_i,\ i = 1,\ldots,n$$

$$\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T,\ b = (b_1,\ldots,b_m) \sim N_m(0, \tau^2 I_m)$$

$$X = \begin{pmatrix} 1 & \pi_1 & \ldots & \pi_1^p \\ . & . & \ldots & . \\ . & . & \ldots & . \\ . & . & \ldots & . \\ 1 & \pi_n & \ldots & \pi_n^p \end{pmatrix},\ Z = \begin{pmatrix} (\pi_1 - k_1)_+^p & \ldots & (\pi_1 - k_m)_+^p \\ . & \ldots & . \\ . & \ldots & . \\ . & \ldots & . \\ (\pi_n - k_1)_+^p & \ldots & (\pi_n - k_m)_+^p \end{pmatrix}.$$

The algorithm of Gibbs sampling for estimating the parameters in Model (3) is as follows:

a) The probit regression model for the binary outcome $y = [y_1, \ldots, y_n]^T$ corresponds to a normal regression model for a latent continuous data $y^* = [y_1^*, \ldots, y_n^*]^T$, which has a truncated multivariate normal distribution with mean $(X\beta + Zb)$ and identity covariance matrix (Albert and Chib 1993), and y_i is the indicator that $y_i^* > 0$. With some initial values of (β, b), values of the latent continuous data y_i^* can be simulated.

b) Specifying a proper flat normal prior distribution $N(0,10^6)$ on β and an inverse gamma distribution $\mathrm{IG}(0.1,0.1)$ on τ^2, the posterior distribution of (β, b, τ^2) given the simulated latent continuous data y^* is

$$\begin{aligned}(\beta,b) \mid \tau^2, y^* &\sim \mathrm{MVN}_{m+p+1}((C^TC + D/\tau^2)^{-1}C^Ty^*, \\ &\qquad (C^TC + D/\tau^2)^{-1}) \\ \tau^2 \mid \beta, b &\sim \mathrm{IG}(0.1 + m/2,\ 0.1 + \|b\|^2/2), \quad (11)\end{aligned}$$

where $C = [X, Z]$ and D is a diagonal matrix with $p+1$ values of 10^{-6} followed by m ones on the diagonal. Gelman (2006) recommended a uniform prior distribution on τ, which results in the posterior distribution for τ^2 as

$$\tau^2 \mid \beta, b \sim \mathrm{IG}((m-1)/2, \|b\|^2/2) \qquad (12)$$

c) At iteration t, draws of $(\beta^{(t)}, b^{(t)}, \tau^{2(t)})$ from the posterior distribution in equation (11) or (12) are used to generate new latent data $\hat{y}^{*(t)}$ conditional on observed binary variable y for the sample, and to obtain the posterior predicted values $\hat{y}^{(t)}$ for non-sample units. We then can obtain draws from the posterior distribution of the finite population proportion at iteration t as

$$\hat{p}_{\mathrm{PR}}^{(t)} = N^{-1}\left(\sum_{i \in s} y_i + \sum_{j \notin s} \hat{y}_j^{(t)}\right)$$

References

Albert, J.H., and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of American Statistical Association*, 88, 669-679.

Basu, D. (1971). An essay on the logical foundations of survey sampling. Part 1, in *Foundations of Statistical Inference*, (Eds., V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart and Winston, 203-242.

Compumine (2007). Re: analysis – Tax audit data mining. Feb. 2007. http://www.compumine.com/web/public/newsletter/20071/tax-audit-data-mining.

Crainiceanu, C.M., Ruppert, D. and Wand, M. (2005). Bayesian analysis for penalized spline regression using WinBUGS. *Journal of Statistical Software*, 14, 2005, 14.

Duchesne, P. (2003). Estimation of a proportion with survey data. *Journal of Statistics Education*, 11, 3.

Eilers, P.H.C., and Marx, B.D. (1996). Flexible smoothing with *B*-splines and penalties (with discussion). *Statistical Science*, 11, 89-121.

Firth, D., and Bennett, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society*, Series B, 60, 3-21.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 3, 515-533.

Hartley, H.O., and Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.

Horvitz, D.G., and Thompson, M.E. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, 47, 663-685.

Isaki, C.T., and Fuller, W.A. (1982). Survey design under a regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Kish, L. (1995). The hundred years' wars of survey sampling. *Statistics in Transition*, 2, 813-830.

Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, 7, 649-673.

Lehtonen, R., and Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology*, 24, 51-55.

Little, R.J.A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99, 546-556.

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

Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G. and Breidt, F.J. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society*, Series B, 70, 265-286.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). *Semiparametric Regression.* Cambridge, UK: Cambridge University Press.

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

Shao, J., and Wu, C.F.J. (1989). A general theory for jackknife variance estimation. *Annuals of Statistics*, 17, 1176-1197.

Smith, T.M.F. (1976). The foundations of survey sampling: A review (with discussion). *Journal of the Royal Statistical Society*, Series A, 139, 183-204.

Smith, T.M.F. (1994). Sample surveys 1975–1990: An age of reconciliation? (with discussion). *International Statistical Review*, 62, 5-34.

Yates, F., and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, Series B, 15, 235-261.

Zheng, H., and Little, R.J.A. (2003). Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19, 99-117.

Zheng, H., and Little, R.J.A. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, 21, 1-20.

Survey Methodology, June 2010
Vol. 36, No. 1, pp. 35-43
Statistics Canada, Catalogue No. 12-001-X

The effect of nonresponse adjustments on variance estimation

David Haziza, Katherine Jenny Thompson and Wesley Yung [1]

Abstract

Many surveys employ weight adjustment procedures to reduce nonresponse bias. These adjustments make use of available auxiliary data. This paper addresses the issue of jackknife variance estimation for estimators that have been adjusted for nonresponse. Using the reverse approach for variance estimation proposed by Fay (1991) and Shao and Steel (1999), we study the effect of not re-calculating the nonresponse weight adjustment within each jackknife replicate. We show that the resulting 'shortcut' jackknife variance estimator tends to overestimate the true variance of point estimators in the case of several weight adjustment procedures used in practice. These theoretical results are confirmed through a simulation study where we compare the shortcut jackknife variance estimator with the full jackknife variance estimator obtained by re-calculating the nonresponse weight adjustment within each jackknife replicate.

Key Words: Calibration; Nonresponse adjustment; Unit nonresponse; Jackknife variance estimator; Linearization variance estimator.

1. Introduction

Unit nonresponse, which occurs when, for a sample unit, all the survey variables are missing or when not enough usable information is available, is unavoidable in surveys. To address this, the nonrespondents are deleted from the data file and the survey weights of the respondents are adjusted to compensate for the deletions. The primary objective of a weight adjustment procedure is to reduce the nonresponse bias, which is introduced when respondents and nonrespondents are different with respect to the survey variables. Key to achieving an efficient bias reduction is the use of powerful auxiliary information available for both respondents and nonrespondents.

In this paper, we consider jackknife variance estimation in the presence of unit nonresponse. This variance estimation method is widely used in practice because of its theoretical properties and computational ease. In contrast to Taylor linearization procedures, the jackknife method does not require a separate derivation for each parameter of interest nor the second-order inclusion probabilities that may be difficult to obtain in complex surveys. When using a jackknife variance estimator in the context of nonresponse, there is some question of whether or not the nonresponse adjustment needs to be replicated (*e.g.*, Valliant 2004). In this paper, we consider two jackknife variance estimators: (i) a *full* jackknife variance estimator which recalculates the nonresponse adjustment factor within each jackknife replicate and (ii) a *shortcut* jackknife variance estimator, which does not. The shortcut jackknife variance estimator is convenient in practice but its theoretical properties were not, to our knowledge, fully studied in the literature. Production reasons tend to drive the usage of a shortcut jackknife variance estimator, since the full jackknife variance estimator in the context of stratified sampling can be quite time-consuming and computer resource-intensive, especially when a survey utilizes a large number of weighting cells. Some recent studies conducted at the U.S. Census Bureau (Thompson 2005 and Ozcoskun, Thompson and Williams 2005) found negligible differences between variance estimates obtained using a fully replicated weight adjustment procedure and those obtained using a "shortcut" procedure with stratified jackknife, delete-a-group jackknife, and modified half sample variance estimators.

Two types of adjustment procedures are commonly used in practice. The first, called *nonresponse propensity weighting* (NPW), consists of first modeling the response propensities and using the inverse of the estimated propensities as the weighting adjustment. The estimated response propensities are typically obtained by fitting a parametric model (*e.g.*, logistic regression model) or by fitting a nonparametric model; *e.g.*, Da Silva and Opsomer (2006). A special case of NPW, which is very popular in practice, consists of first dividing the respondents and nonrespondents into weighting classes and adjusting the design weights of respondents by the inverse of the response rate within each class. These classes are formed on the basis of auxiliary information recorded for all units in the sample; see, for example, Eltinge and Yansaneh (1997) and Little (1986). The second type of adjustment procedures, called *nonresponse calibration weighting* (NCW) can be seen as an extension of the calibration approach (Deville and Särndal 1992) adapted to the context of unit nonresponse. The reader is referred to Särndal and Lundström (2005), Kott (2006) and Brick and Montaquila (2008) for a comprehensive overview of NPW and NWC. In some

1. David Haziza, Département de mathématiques et de statistique, Université de Montréal, Montréal, H3C 3J7, Canada. E-mail: David.haziza@umontreal.ca; Katherine Jenny Thompson, U.S. Census Bureau, Washington, DC 20233. E-mail: Katherine.J.Thompson@census.gov; Wesley Yung, Statistics Canada, Ottawa, Ontario, K1A 0T6. E-mail: wesley.yung@statcan.gc.ca.

situations, NPW and NCW lead to the same estimator; for example, the count-adjusted estimator presented below (see expression (1.4)). In this paper, we focus on NCW. The problem of variance estimation in the context of NPW has been recently studied by Kim and Kim (2007).

Consider a finite population U of size N. The objective is to estimate the population total $Y = \sum_{i \in U} y_i$, of a variable of interest y. Suppose that a random sample s of size n is selected from U according to a given design $p(s)$. In the case of complete data, a basic estimator of Y is the well-known expansion estimator given by

$$\hat{Y}_\pi = \sum_{i \in s} d_i y_i \tag{1.1}$$

where $d_i = 1/\pi_i$ denotes the design weight attached to unit i and $\pi_i = P(i \in s)$ denotes its first-order probability of inclusion in the sample. In the presence of unit nonresponse, only a subset of s is observed, and so the computation of $\hat{Y}_\pi$ in (1.1) is not possible.

To define a nonresponse adjusted estimator of Y, we assume that a vector of auxiliary variables $\mathbf{x}$ is available for all the sampled units (respondents and nonrespondents) so that the vector of estimated totals, $\hat{\mathbf{X}}_\pi = \sum_{i \in s} d_i \mathbf{x}_i$, is available. We also assume that a vector of instrumental variables $\mathbf{z}$, of the same dimension as $\mathbf{x}$, is available for the respondents. Let r_i be a response indicator attached to unit i such that $r_i = 1$ if unit i is a responding unit and $r_i = 0$, otherwise. To estimate Y, we consider calibration estimators of the form

$$\hat{Y}_{\text{CAL}} = \sum_{i \in s} w_i r_i y_i, \tag{1.2}$$

where $w_i = d_i g_i$ and g_i is a nonresponse weighting adjustment factor attached to unit i and given by

$$g_i = 1 + (\hat{\mathbf{X}}_\pi - \hat{\mathbf{X}}_r)' \hat{\mathbf{T}}_r^{-1} \mathbf{z}_i, \tag{1.3}$$

where $\hat{\mathbf{X}}_r = \sum_{i \in s} d_i r_i \mathbf{x}_i$ and $\hat{\mathbf{T}}_r = \sum_{i \in s} d_i r_i \mathbf{z}_i \mathbf{x}_i'$. When $\mathbf{z}_i = \mathbf{x}_i/\nu_i$, where ν_i is a known constant, then the estimator (1.3) is identical to the *InfoS* estimator given in Särndal and Lundström (2005, equation 7.15). The properties of the estimator (1.2) were studied by Deville (2002), Sautory (2003), Särndal and Lundström (2005) and Kott (2006), among others.

In this paper, the properties (*e.g.*, bias and variance) of $\hat{Y}_{\text{CAL}}$ are studied using the nonresponse model (NM) approach, under which inference is made with respect to the joint distribution induced by the sampling design and the nonresponse mechanism, $q(\mathbf{r} \mid \mathbf{I})$, where $\mathbf{I} = (I_1, ..., I_N)'$ is the vector of sample selection indicators such that $I_i = 1$ if unit i is selected in the sample and $I_i = 0$, otherwise and $\mathbf{r} = (r_1, ..., r_N)'$ is the vector of response indicators. Let $p_i = P(r_i = 1 \mid \mathbf{I}, I_i = 1)$ be the response probability for unit i. We assume that $p_i > 0$ for all i and that the units respond independently of one another; that is, $p_{ij} = P(r_i = 1, r_j = 1 \mid \mathbf{I}, I_i = 1, I_j = 1, i \neq j) = p_i p_j$.

The estimator $\hat{Y}_{\text{CAL}}$ is asymptotically unbiased for the true total Y if (i) $p_i^{-1} = 1 + \boldsymbol{\lambda}' \mathbf{z}_i$ for all $i \in U$, where $\boldsymbol{\lambda}$ is a vector of unknown constants or (ii) $y_i = \mathbf{x}_i' \boldsymbol{\beta}$ for all $i \in U$, where $\boldsymbol{\beta}$ is a vector of constants; see Särndal and Lundström (2005, chapter 9.5). If the condition (i) is satisfied, the point estimator $\hat{Y}_{\text{CAL}}$ is asymptotically unbiased for Y regardless of the variable of interest y being estimated. Also, it follows from (ii) that $\hat{Y}_{\text{CAL}}$ has a small bias if the residuals $E_i = y_i - \mathbf{x}_i' \mathbf{B}$, are small, where $\mathbf{B} = (\sum_{i \in U} \mathbf{z}_i \mathbf{x}_i')^{-1} \sum_{i \in U} \mathbf{z}_i y_i$. Therefore, the bias of the estimator $\hat{Y}_{\text{CAL}}$ is small if the vector $\mathbf{x}$ explains the variable of interest y. In the case of several variables of interest, note that the vector $\mathbf{x}$ may explain a given variable of interest well but may not be related to all, in which case some estimates could be potentially biased. We assume that $\hat{Y}_{\text{CAL}}$ is asymptotically unbiased for Y, so that the bias of the estimators under consideration is not an issue in the reminder of the paper.

We consider three special cases of (1.2) that are of interest in practice (see also Kalton and Flores-Cervantes 2003). First, let $\boldsymbol{\delta} = (\delta_{i1}, ..., \delta_{ic}, ..., \delta_{iC})'$ be a C-vector of weighting class indicators attached to unit i such that $\delta_{ic} = 1$ if unit i belongs to class c and $\delta_{ic} = 0$, otherwise for $c = 1, ..., C$. If $\mathbf{x}_i = \mathbf{z}_i = \boldsymbol{\delta}_i$, the adjustment factor g_i given by (1.3) reduces to $g_i = \hat{N}_c / \hat{N}_{rc}\, \delta_{ic}$, where $\hat{N}_c = \sum_{i \in s} d_i \delta_{ic}$ and $\hat{N}_{rc} = \sum_{i \in s} d_i r_i \delta_{ic}$. That is, the nonresponse weighting adjustment factor for a weighting cell is calculated as the sample-weighted number of sampled units in the weighting cell divided by the sample-weighted number of responding units in the weighting cell. We refer to this weight adjustment procedure as the *count adjustment* procedure. It follows that the estimator (1.2) reduces to the count adjusted estimator

$$\hat{Y}_{\text{count}} = \sum_{c=1}^{C} \frac{\hat{N}_c}{\hat{N}_{rc}} \hat{Y}_{rc}, \tag{1.4}$$

where

$$\hat{Y}_{rc} = \sum_{i \in s} d_i r_i \delta_{ic} y_i.$$

The second special case of (1.2) assumes that a continuous variable x is available for all the sampled units. Let $\mathbf{x}_i = (\delta_{i1} x_i, ..., \delta_{ic} x_i, ..., \delta_{iC} x_i)'$ and $\mathbf{z}_i = \boldsymbol{\delta}_i$. In this case, the adjustment factor g_i given by (1.3) reduces to $g_i = \hat{X}_c / \hat{X}_{rc}\, \delta_{ic}$ if unit i belongs to class c, where $\hat{X}_c = \sum_{i \in s} d_i \delta_{ic} x_i$ and $\hat{X}_{rc} = \sum_{i \in s} d_i r_i \delta_{ic} x_i$. Here, the nonresponse weighting adjustment factor for a weighting class c is the sum of the sample-weighted auxiliary data for units in the weighting cell divided by the sum of the

sample-weighted auxiliary data for all responding units in the weighting cell. We refer to this weight adjustment procedure as the *ratio adjustment* procedure. The estimator (1.2) reduces to the ratio adjusted estimator

$$\hat{Y}_{\text{ratio}} = \sum_{c=1}^{C} \frac{\hat{X}_c}{\hat{X}_{rc}} \hat{Y}_{rc}. \qquad (1.5)$$

Note that the count adjusted estimator (1.4) is a special case of the ratio adjusted estimator when $x_i = 1$ for all the sampled population units.

Finally, if $\mathbf{x}_i = \mathbf{z}_i = (\delta_{i1}, ..., \delta_{ic}, ..., \delta_{iC}, \delta_{i1} x_i, ..., \delta_{ic} x_i, ..., \delta_{iC} x_i)'$, we obtain another special case of (1.2). In this case, the adjustment factor g_i given by (1.3) reduces to

$$g_i = \hat{N}_c \left[1 + (\bar{x}_c - \bar{x}_{rc}) \frac{(x_i - \bar{x}_{rc})}{\sum_{i \in s} r_i \delta_{ic} (x_i - \bar{x}_{rc})^2} \right],$$

if unit i belongs to class c, where $\bar{x}_c = \hat{X}_c / \hat{N}_c$ and $\bar{x}_{rc} = \hat{X}_{rc} / \hat{N}_{rc}$. We refer to this weight adjustment procedure as the *simple linear regression adjustment* procedure. The estimator (1.2) reduces to the simple linear regression adjusted estimator

$$\hat{Y}_{slreg} = \sum_{c=1}^{C} \hat{N}_c [\hat{Y}_{rc} + (\bar{x}_c - \bar{x}_{rc}) \hat{B}_{rc}], \qquad (1.6)$$

where

$$\hat{B}_{rc} = \frac{\sum_{i \in s} d_i r_i \delta_{ci} (x_i - \bar{x}_{rc})(y_i - \bar{y}_{rc})}{\sum_{i \in s} d_i r_i \delta_{ic} (x_i - \bar{x}_{rc})^2}.$$

The estimators (1.4)-(1.6) use some form of weighting adjustment within classes. All of them are asymptotically unbiased for Y if the units have equal response probabilities within classes (*i.e.*, a uniform nonresponse mechanism within classes). This condition is a special case of condition (i) discussed above.

In this paper, we show that the shortcut jackknife variance estimator that treats the adjustment factors as fixed, tends to overestimate the true variance of $\hat{Y}_{\text{CAL}}$, at least in some simple cases. We build on earlier research by Thompson and Yung (2006) who derived expressions of the linearization version for both the full and shortcut jackknife variance estimators and evaluated these expressions empirically using data from the Annual Capital Expenditures Survey (ACES), conducted at the U.S. Census Bureau. In the context of NPW, it is interesting to note that Kim and Kim (2007) showed that treating the estimated response probabilities as fixed leads to an overestimation of the true variance when the sampling weights are not used in estimating these probabilities. Beaumont (2005) obtained similar results in the context of imputation when the response probabilities are estimated using a logistic regression model.

In Section 2, we discuss the full and shortcut jackknife variance estimators and show that the shortcut estimator is asymptotically biased. The severity of this bias is evaluated for two commonly used sample designs in Section 3. Section 4 presents the results of a simulation study comparing the full and shortcut jackknife variance estimators. We conclude in Section 5 with some general observations.

2. Jackknife variance estimation

Traditionally, variance estimation in the context of nonresponse has been performed using the two-phase framework, which consists of viewing nonresponse as a second-phase of selection. Instead, we consider the reverse framework that was proposed by Fay (1991) and further developed by Shao and Steel (1999). This framework provides a theoretical basis for studying the properties of jackknife variance estimators and can be described as follows: first, applying the nonresponse mechanism, the population U is randomly divided into a population of respondents U_r and a population of nonrespondents U_m. Then, given (U_r, U_m), the random sample s is selected according to the chosen sampling design. The total variance of $\hat{Y}_{\text{CAL}}$ can be expressed as

$$V(\hat{Y}_{\text{CAL}}) = E_q V_p (\hat{Y}_{\text{CAL}} \mid \mathbf{r}) + V_q E_p (\hat{Y}_{\text{CAL}} \mid \mathbf{r}), \qquad (2.1)$$

where $E_p(.)$ and $V_p(.)$ denote the expectation and the variance with respect to the sampling design and $E_q(.)$ and $V_q(.)$ denote the expectation and variance with respect to the nonresponse mechanism, $q(\mathbf{r} \mid \mathbf{I})$.

In this section, we focus on stratified simple random sampling, which is the design typically used in business surveys. With this sample design, the population U is partitioned into L strata $U_1, ..., U_L$ of size $N_1, ..., N_L$, respectively. A simple random sample without replacement s_h, of size n_h, is selected from stratum h, $h = 1, ..., L$. Each within-stratum sample is selected independently, and we assume that $n_h \geq 2$ for all h. In this context, the design weight of unit i in stratum h is $d_{hi} = N_h / n_h$. A full jackknife variance estimator of $\hat{Y}_{\text{CAL}}$, under stratified simple random sampling, is obtained as follows:

(i) remove unit (gj) from the sample, $g = 1, ..., L$; $j = 1, ..., n_h$;

(ii) adjust the design weights d_{hi} to obtain the jackknife weights $d_{hi(gj)}$, where $d_{hi(gj)}$ is given by

$$d_{hi(gj)} = \begin{cases} 0 & \text{if } (hi) = (gj) \\ \dfrac{n_g}{n_g - 1} d_{gi} & \text{if } h = g,\ i \neq j \\ d_{hi} & \text{otherwise} \end{cases}$$

(iii) compute the estimator $\hat{Y}_{\mathrm{CAL}(gj)}$ in the same way as $\hat{Y}_{\mathrm{CAL}}$ with the jackknife weights $d_{hi(gj)}$ instead of the design weights d_{hi}; that is, $\hat{Y}_{\mathrm{CAL}(gj)} = \sum_{(hi)\in s} w_{hi(gj)} r_{hi} y_{hi}$, where $w_{hi(gj)} = d_{hi(gj)} g_{hi(gj)}$ with $g_{hi(gj)} = 1 + (\hat{\mathbf{X}}_{\pi(gj)} - \hat{\mathbf{X}}_{r(gj)})' \hat{\mathbf{T}}_{r(gj)}^{-1} \mathbf{z}_{hi}$, $\hat{\mathbf{X}}_{\pi(gj)} = \sum_{i\in s} d_{hi(gj)} \mathbf{x}_{hi}$, $\hat{\mathbf{X}}_{r(gj)} = \sum_{(hi)\in s} d_{hi(gj)} r_{hi} \mathbf{x}_{hi}$ and $\hat{\mathbf{T}}_{r(gj)} = \sum_{(hi)\in s} d_{hi(gj)} r_{hi} \mathbf{z}_{hi} \mathbf{x}_{hi}'$.

(iv) replace the unit deleted in step (i) back into the sample;

(v) repeat steps (i)-(iv) for all (gj) units, $g = 1, ..., L$; $j = 1, ..., n_h$.

Note that the nonresponse adjustment factors g_{hi} are recalculated in each replicate. This leads to the full jackknife variance estimator

$$v_{JF} = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j\in s_h} (\hat{Y}_{\mathrm{CAL}(gj)} - \hat{Y}_{\mathrm{CAL}})^2. \qquad (2.2)$$

The variance estimator v_{JF} is an estimator of the first term on the right hand side of (2.1), $E_q V_p(\hat{Y}_{\mathrm{CAL}} \mid \mathbf{r})$. This term represents the design variance that we would have obtained if the responding units were selected using stratified simple random sampling with replacement, or equivalently, if the stratum sampling fractions, (n_h / N_h) are negligible. In other words, the full jackknife variance estimator (2.2) is an estimator of the sampling variance conditional on the vector of response indicators $\mathbf{r}$. Therefore, v_{JF} is asymptotically unbiased and consistent for $E_q V_p(\hat{Y}_{\mathrm{CAL}} \mid \mathbf{r})$ under stratified simple random sampling with replacement sampling regardless of the validity of the underlying assumptions. Note that since v_{JF} is an estimator of a sampling variance, it can be readily obtained using software designed for complete-data jackknife variance estimation. In other words, no specialized software is needed. Also, note that the second term on the right hand side of (2.1), $V_q E_p(\hat{Y}_{\mathrm{CAL}} \mid \mathbf{r})$, is not accounted for. Thus, the full jackknife variance estimator does not track the second term in (2.1). However, the contribution of this term to the total variance is negligible if the stratum sampling fractions, n_h / N_h, are negligible. As a result, v_{JF} is asymptotically unbiased and consistent for the total variance, $V(\hat{Y}_{\mathrm{CAL}})$. That is, $E_{pq}(v_{JF}) \approx V(\hat{Y}_{\mathrm{CAL}})$. Since the goal of the research is to compare the full and shortcut jackknife estimators, in the remainder of the paper, we assume that the stratum sampling fractions are negligible and focus on estimates of totals, so that we can omit the estimation of the second term in (2.1). We note that even if the second term is not negligible, our comparisons are valid as both the full jackknife and shortcut estimators would underestimate the total variance by the same term.

A shortcut jackknife variance estimator of $\hat{Y}_{\mathrm{CAL}}$ is given by

$$v_{JS} = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j\in s_h} (\hat{Y}^*_{\mathrm{CAL}(gj)} - \hat{Y}_{\mathrm{CAL}})^2, \qquad (2.3)$$

where $\hat{Y}^*_{\mathrm{CAL}(gj)} = \sum_{(hi)\in s} d_{hi(gj)} g_{hi} r_{hi} y_{hi}$. Note that the nonresponse weighting adjustment factors g_{hi} are not recalculated in each jackknife replicate. In other words, the factors g_{hi} are treated as constants, which is inappropriate since they depend on the sample and the set of respondents. Therefore, we have $E_{pq}(v_{JS}) \neq V(\hat{Y}_{\mathrm{CAL}})$, in general, and the shortcut variance estimator, v_{JS}, is biased.

To study the magnitude of the bias of v_{JS}, we consider the difference of the two jackknife variance estimators, $D = v_{JS} - v_{JF}$. Since the variance estimator v_{JF} is an asymptotically unbiased estimator of the term $V_p(\hat{Y}_{\mathrm{CAL}} \mid \mathbf{r})$, it is asymptotically equivalent to a variance estimator obtained using a first-order Taylor expansion. The resulting variance estimator, denoted by $\tilde{v}_{JF}$, is the linearization jackknife variance estimator studied by Yung and Rao (2000). Similarly, the shortcut jackknife variance estimator v_{JS} is asymptotically equivalent to a variance estimator of $V_p(\hat{Y}_{\mathrm{CAL}} \mid \mathbf{r})$ obtained by treating the nonresponse weighting adjustment factors g_{hi} as constants. We denote this variance estimator by $\tilde{v}_{JS}$. The quantity D can thus be approximated by $\tilde{D} = \tilde{v}_{JS} - \tilde{v}_{JF}$. For this approximation to be valid, we assume the number of respondents to be large.

Noting that $\mathrm{Bias}(v_{JF}) = E_{pq}(v_{JF}) - V(\hat{Y}_{\mathrm{CAL}}) \approx 0$, it follows that the bias of v_{JS}, $\mathrm{Bias}(v_{JS}) = E_{pq}(v_{JS}) - V(\hat{Y}_{\mathrm{CAL}})$, can be approximated by $E_{pq}(D) \approx E_{pq}(\tilde{D})$. Let $v(y)$ denote the variance estimator of the complete data estimator (1.1). Using a first-order Taylor expansion, it can be shown that an estimator of $V_p(\hat{Y}_{\mathrm{CAL}} \mid \mathbf{r})$ is given by

$$\tilde{v}_{JF} = v(\hat{\xi}) \qquad (2.4)$$

where

$$\hat{\xi}_{hi} = \mathbf{x}_{hi}' \hat{\mathbf{B}}_r + g_{hi} r_{hi} e_{hi},$$

with $e_{hi} = (y_{hi} - \mathbf{x}_{hi}' \hat{\mathbf{B}}_r)$ and $\hat{\mathbf{B}}_r = \hat{\mathbf{T}}_r^{-1} \sum_{(hi)\in s} d_{hi} r_{hi} \mathbf{z}_{hi} y_{hi}$. On the other hand, treating the g_{hi}'s as constants implies that $\hat{Y}_{\mathrm{CAL}}$ is linear in the design weights d_{hi}. It follows that $\tilde{v}_{JS}$ is given by

$$\tilde{v}_{JS} = v(\psi), \qquad (2.5)$$

where $\psi_{hi} = g_{hi} r_{hi} y_{hi}$.

For example, for either a fixed size or a random size sampling design, a possible variance estimator is

$$\tilde{v}_{JF} = \sum_{i\in s} \sum_{j\in s} \Delta_{ij} \hat{\xi}_i \hat{\xi}_j,$$

where $\Delta_{ij} = (\pi_{ij} - \pi_i \pi_j) / \pi_{ij} \pi_i \pi_j$ and π_{ij} is the second-order inclusion probability of units i and j. Note that $\pi_{ii} = \pi_i$. Similarly, we have

$$\tilde{v}_{JS} = \sum_{i\in s}\sum_{j\in s}\Delta_{ij}\,\psi_i\,\psi_j.$$

3. Bias of v_{JS} in some special cases

3.1 Simple random sampling without replacement

In this section, we assume that the sample s has been selected according to simple random sampling without replacement. We also assume that the sampling fraction n/N is negligible and that the number of respondents r is large. Finally, we assume a single weighting class. Although the above situation is not realistic in practice, it provides some insight into the asymptotic bias of v_{JS}.

In the case of the ratio adjusted estimator (1.5), we can show that $\tilde{D}$ is approximately given by

$$\tilde{D} = \frac{N^2}{r}\left(1-\frac{r}{n}\right)\left\{\left(\frac{\bar{x}}{\bar{x}_r}\right)^2 (s_{yr}^2 - s_{er}^2) + 2\left(\frac{\bar{x}}{\bar{x}_r}\right)\hat{R}_r\left[\left(\frac{\bar{x}}{\bar{x}_r}\right)-1\right]\frac{s_{exr}}{n} + \hat{R}_r^2\left[\left(\frac{\bar{x}}{\bar{x}_r}\right)^2 s_{xr}^2 - s_x^2\right] + \left(\frac{\bar{x}}{\bar{x}_r}\right)^2 \bar{y}_r^2\right\}, \quad (3.1)$$

where $(\bar{x}_r, \bar{y}_r) = 1/r\sum_{i\in s} r_i (x_i, y_i)$ denote the mean of the respondents for variable x and y respectively and r is the number of respondents, $\hat{R}_r = \bar{y}_r/\bar{x}_r$, $s_{xr}^2 = 1/(r-1)\sum_{i\in s} r_i (x_i - \bar{x}_r)^2$, $s_x^2 = 1/(n-1)\sum_{i\in s}(x_i - \bar{x})^2$ with $\bar{x} = 1/n\sum_{i\in s} x_i$, $s_{er}^2 = 1/(r-1)\sum_{i\in s} r_i (y_i - \hat{R}_r x_i)^2$ and $s_{exr} = 1/(r-1)\sum_{i\in s} r_i (y_i - \hat{R}_r x_i) x_i$. If we further assume that all units have equal response probabilities (*i.e.*, a uniform response mechanism), we have $\bar{x}/\bar{x}_r \xrightarrow{p} 1$ and $s_{xr}^2/s_x^2 \xrightarrow{p} 1$. In this case, the asymptotic bias of v_{JS} is given by

$$\text{Bias}(v_{JS}) \approx E_{pq}(\tilde{D}) \approx \frac{N^2}{E_{pq}(r)}\left(1 - E_{pq}\left(\frac{r}{n}\right)\right) S_y^2\left(\frac{1}{\text{CV}(y)^2} + 2\frac{\text{CV}(x)}{\text{CV}(y)}\rho_{xy} - \frac{\text{CV}(x)^2}{\text{CV}(y)^2}\right), \quad (3.2)$$

where $\text{CV}(x) = S_x/\bar{X}$ and $\text{CV}(y) = S_y/\bar{Y}$ denote the population coefficients of variation for variables x and y, respectively with $S_y^2 = 1/(N-1)\sum_{i\in U}(y_i - \bar{Y})^2$ and $\bar{Y} = 1/N\sum_{i\in U} y_i$, S_x^2 and $\bar{X}$ are defined similarly, and ρ_{xy} denotes the finite population coefficient of correlation for variables x and y. From (3.2), it follows that the asymptotic bias of v_{JS} is nonnegative if and only if

$$B_0 < \frac{\bar{Y}}{2}\left(\frac{1+\text{CV}(x)^2}{\text{CV}(x)^2}\right), \quad (3.3)$$

provided $0 < E_{pq}(r/n) < 1$, where $B_0 = \bar{Y} - B_1\bar{X}$ is the finite population intercept of the least squares line when regressing y on x with

$$B_1 = \frac{\sum_{i\in U}(x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i\in U}(x_i - \bar{X})^2}.$$

From (3.2), it is clear that the bias of v_{JS} increases if (i) the expected response rate $E_{pq}(r/n)$ decreases; (ii) ρ_{xy} increases; (iii) $\text{CV}(y)$ decreases; or (iv) $\text{CV}(x)$ increases. Also, it follows from (3.3) that v_{JS} overestimates the true variance when the intercept B_0 is not too large. Table 1 illustrates the relationship between $\text{CV}(x)$ and the condition in (3.3). For example, when $\text{CV}(x) = 0$, v_{JS} always overestimates the true variance since, in this case, the condition (3.3) reduces to $B_0 < \infty$, which is always satisfied. This result is not surprising because when $\text{CV}(x) = 0$, the x-values are all equal and the ratio adjusted estimator (1.5) is identical to the count adjusted estimator (1.4). As we discuss below, v_{JS} always overestimates the true variance in this case. When $\text{CV}(x)$ is large (*e.g.*, $\text{CV}(x) = 2$), v_{JS} overestimates the true variance if and only if $B_0 < 0.625\bar{Y}$ The latter condition is satisfied if the intercept is not "too far" from the origin. Therefore, if the relationship between y and x goes through the origin (*i.e.*, if the ratio model holds), the shortcut variance estimator will overestimate the true variance. However, if the ratio adjusted estimator is used when the ratio model does not hold, such as when $B_0 \geq 0.625\bar{Y}$, the shortcut variance estimator v_{JS} will underestimate the true variance. In conclusion, we can expect v_{JS} to overestimate the true variance when a ratio adjustment procedure is used unless the ratio model is highly misspecified for the data at hand, which could happen, for example, if the variables y and x are negatively correlated.

Table 1
Relationship between $\text{CV}(x)$ and the condition in (3.3)

$\text{CV}(x)$	$\frac{\bar{Y}}{2}\left(\frac{1+\text{CV}(x)^2}{\text{CV}(x)^2}\right)$
0	∞
0.1	$50.5\,\bar{Y}$
0.5	$2.5\,\bar{Y}$
1	$2\,\bar{Y}$
1.5	$0.722\,\bar{Y}$
2	$0.625\,\bar{Y}$

Turning to the count adjusted estimator (1.4), we let $x_i = 1$ for all i in (3.1) and obtain

$$\tilde{D} = \frac{N^2}{r}\left(1-\frac{r}{n}\right)\bar{y}_r^2. \quad (3.4)$$

It follows from (3.4) that the relative bias of v_{JS}, $\mathrm{RB}(v_{JS}) = \mathrm{Bias}(v_{JS})/V(\hat{Y}_{\mathrm{CAL}})$, can be approximated by by $E_{pq}(R\tilde{D})$ where $R\tilde{D} = \tilde{D}/\tilde{v}_{JF}$. Under a uniform nonresponse mechanism, straightforward algebra leads to

$$\mathrm{RB}(v_{JS}) \approx E_{pq}(R\tilde{D}) \approx \left(1 - E_{pq}\left(\frac{r}{n}\right)\right)\frac{1}{\mathrm{CV}(y)^2}. \quad (3.5)$$

The expression (3.5) shows that, in the case of the count adjusted estimator (1.4), v_{JS} always overestimates the true variance. The magnitude of the overestimation increases as the expected response rate $E_{pq}(r/n)$ decreases or when $\mathrm{CV}(y)$ decreases. For example, if the expected response rate is equal to 70% and $\mathrm{CV}(y) = 1$, we have $E_{pq}(R\tilde{D}) = 1.3$ so the shortcut jackknife variance estimator, v_{JS}, is on average 30% larger than the true variance of $\hat{Y}_{\mathrm{CAL}}$. On the other hand, if the response rate is equal to 70% and $\mathrm{CV}(y) = 0.5$, we have $E_{pq}(R\tilde{D}) = 5.3$, in which case the overestimation is considerable.

Finally, we turn to the case of the simple linear regression adjusted estimator (1.6). Under a uniform nonresponse mechanism, it can be shown that the asymptotic bias of v_{JS} is given by

$$\begin{aligned}\mathrm{Bias}(v_{JS}) &\approx E_{pq}(\tilde{D}) \\ &\approx \frac{N^2}{E_{pq}(r)}\left(1 - E_{pq}\left(\frac{r}{n}\right)\right) \\ &\quad S_y^2\left(\frac{1}{\mathrm{CV}(y)^2} + \rho_{xy}^2\right) \geq 0. \quad (3.6)\end{aligned}$$

From (3.6), it follows that v_{JS} always overestimates the true variance in the case of the simple linear regression adjusted estimator (1.6). The bias (3.6) increases if (i) the expected response rate decreases; (ii) ρ_{xy}^2 increases; or (iii) $\mathrm{CV}(y)$ decreases.

3.2 Stratified simple random sampling: Weighting classes are identical to strata

In this section, we assume that the weighting classes coincide with the original design strata. This situation is not uncommon in practice, especially in business surveys. If the strata are such that the units within stratum have approximately equal response propensities (*i.e.*, uniform response within stratum), expressions for the bias of v_{JS} are readily obtained from expressions (3.2), (3.4) and (3.6).

For the ratio adjusted estimator, expression (3.2) can be readily extended to the case of stratified simple random sampling to obtain

$$\begin{aligned}\mathrm{Bias}(v_{JS}) &\approx E_{pq}(\tilde{D}) \\ &\approx \sum_{h=1}^{L}\frac{N_h^2}{E_{pq}(r_h)}\left(1 - E_{pq}\left(\frac{r_h}{n_h}\right)\right) \\ &\quad S_{yh}^2\left(\frac{1}{\mathrm{CV}_h(y)^2} + 2\frac{\mathrm{CV}_h(x)}{\mathrm{CV}_h(y)}\rho_{hxy} - \frac{\mathrm{CV}_h(x)^2}{\mathrm{CV}_h(y)^2}\right), \quad (3.7)\end{aligned}$$

where the quantities r_h, $\mathrm{CV}_h(x)$, $\mathrm{CV}_h(y)$, S_{yh}^2 and ρ_{hxy} correspond to r, $\mathrm{CV}(x)$, $\mathrm{CV}(y)$, S_y^2 and ρ_{xy} computed in each stratum.

For the count adjusted estimator, expression (3.4) can be readily extended to the case of stratified simple random sampling to obtain

$$\begin{aligned}\mathrm{Bias}(v_{JS}) &\approx E_{pq}(\tilde{D}) \\ &\approx \sum_{h=1}^{L}\frac{N_h^2}{E_{pq}(r_h)}\left(1 - E_{pq}\left(\frac{r_h}{n_h}\right)\right)\frac{S_{yh}^2}{\mathrm{CV}_h(y)^2}. \quad (3.8)\end{aligned}$$

Finally, for the simple linear regression adjusted estimator, expression (3.6) can be readily extended to the case of stratified simple random sampling to obtain

$$\begin{aligned}\mathrm{Bias}(v_{JS}) &\approx E_{pq}(\tilde{D}) \\ &\approx \sum_{h=1}^{L}\frac{N_h^2}{E_{pq}(r_h)}\left(1 - E_{pq}\left(\frac{r_h}{n_h}\right)\right) \\ &\quad S_{yh}^2\left(\frac{1}{\mathrm{CV}_h(y)^2} + \rho_{hxy}^2\right). \quad (3.9)\end{aligned}$$

From the expressions (3.7)-(3.9), it follows that the use of the shortcut jackknife variance estimator requires some caution. Indeed, even if the bias of the shortcut jackknife variance estimator is small in each stratum, they might sum up to a considerable bias at the population level if the biases are in the same direction.

4. Simulation study

A simulation study was performed to compare the statistical properties of the shortcut and the full jackknife variance estimators under varying conditions. Five different stratified populations of 30,000 units each with two variables were generated. First, the *x*-values were generated from a Gamma distribution with parameters α and λ. Then given the *x*-values, the *y*-values were generated according to the following model:

$$y_{hi} = \beta_0 + \beta_1 x_{hi} + \varepsilon_{hi},$$

where $\varepsilon_{hi} \sim N(0, \sigma_{\varepsilon h}^2)$. The variance and $\sigma_{\varepsilon h}^2$ was set such that the coefficient of correlation (denoted ρ_{xy}) between x_{hi} and y_{hi} is equal to 0.7 in all the populations. Each population was stratified into three strata, each with 10,000 units. The parameters of the simulated populations appear in Table 2.

Population 1 fits the ratio model very well with an intercept of zero in all strata. Population 2 has a non-negligible intercept term in all three strata. Population 3 is a mix of populations 1 and 2, where the ratio model fits well for strata 2 and 3 but not for stratum 1. Population 4 is

similar to population 1 except units in strata 1 and 2 have a 70% chance of reporting a zero. This population is intended to mimic the situation of the Annual Capital Expenditures Survey (ACES) of the U.S. Census Bureau, which provided the motivation for this research. The ACES employs a shortcut jackknife variance estimator that, empirically, has been shown to be close to the full jackknife variance estimates. Its population is characterized with many zeros for capital expenditures in the majority of sampled small and medium businesses, with the majority of the reported expenditures being provided by large businesses. Population 5 was generated to show that the shortcut estimator for the ratio adjusted estimator can actually have a negative bias when the ratio model is misspecified (demonstrated in expression (3.3) for a simple random sample). For this population, the intercept term is highly significant in all strata.

Table 2
Population parameters

Population	β_0 (Within Stratum)			β_1 (Within Stratum)			α	λ	CV(x)	CV(y)
	1	2	3	1	2	3				
1	0	0	0	2	4	6	4	5	50%	76%
2	120	240	360	2	4	6	4	5	50%	44%
3	120	0	0	2	4	6	4	5	50%	51%
4	0	0	0	2	4	6	4	5	50%	134%
5	50	200	300	0.5	1	2	4	5	200%	63%

From each population, 5,000 stratified simple random samples of size 300 (100 units per stratum) were drawn. In each sample, nonresponse was generated using a uniform response mechanism within each stratum with probabilities of response equal to 60% in stratum 1, 70% in stratum 2 and 90% in stratum 3. This response pattern is not uncommon in business surveys where more follow-up is performed for the medium and large size units (strata 2 and 3).

In each sample, both the count adjusted and the ratio adjusted estimators, given respectively by (1.4) and (1.5), were calculated using the strata as weighting classes. The variance of the point estimators was estimated by v_{JF} and v_{JS}, given respectively by (2.2) and (2.3). As a measure of the bias of a variance estimator v, we used the Monte Carlo percent relative bias given by

$$\mathrm{RB}_{\mathrm{MC}}(v) = \frac{1}{5{,}000}\sum_{t=1}^{5{,}000}\frac{v^{(t)} - \mathrm{MSE}_{\mathrm{MC}}(\hat{Y}_{\mathrm{CAL}})}{\mathrm{MSE}_{\mathrm{MC}}(\hat{Y}_{\mathrm{CAL}})}\times 100,$$

where $v^{(t)}$ is the variance estimate obtained from the t^{th} sample, and $\mathrm{MSE}_{\mathrm{MC}}(\hat{Y}_{\mathrm{CAL}})$ is the Monte Carlo Mean Squared Error (MSE) defined by

$$\mathrm{MSE}_{\mathrm{MC}}(\hat{Y}_{\mathrm{CAL}}) = \frac{1}{50{,}000}\sum_{t=1}^{50{,}000}(\hat{Y}_{\mathrm{CAL}}^{(t)} - Y)^2,$$

where $\hat{Y}_{\mathrm{CAL}}^{(t)}$ is the (ratio or count adjusted) estimate of Y for the t^{th} sample. Table 3 shows the Monte Carlo percent relative bias for both the count adjusted and the ratio adjusted estimators.

Table 3
Monte Carlo percent relative bias for the shortcut and full jackknife variance estimators

Population	Count adjusted estimator		Ratio adjusted estimator	
	$\mathrm{RB}_{\mathrm{MC}}(v_{JS})$	$\mathrm{RB}_{\mathrm{MC}}(v_{JF})$	$\mathrm{RB}_{\mathrm{MC}}(v_{JS})$	$\mathrm{RB}_{\mathrm{MC}}(v_{JF})$
1	57.3%	1.1%	80.5%	-0.3%
2	877.1%	0.4%	364.7%	0.5%
3	220.7%	0.6%	185.9%	-0.2%
4	21.6%	0.6%	29.1%	1.4%
5	266.4%	0.2%	-67.2%	5.0%

As expected, the shortcut estimator overestimates the Monte Carlo MSE for the count adjusted estimator for all populations. The overestimation varies from approximately 20% in population 4 to over 800% in population 2. From expression (3.8), we see that the bias of v_{JS} depends on the response rate and $\bar{y}_h^2$. Population 2 has a large intercept term which increases $\mathrm{CV}_h(y)$ in all strata, which in turn increases the bias of v_{JS}. Population 3 is similar to population 2 except only the first stratum has a large intercept term. As expected, the bias of v_{JS} in this population is between those of populations 1 and 2. Population 4 is the one generated to mimic the ACES population with some units' values replaced by zero in strata 1 and 2. The Monte Carlo relative bias of 21.6% is, for the most part, coming from the third stratum where no units have been replaced with zero (this can be seen using expression (3.8)). In comparison, for all five populations the full jackknife variance estimator is tracking the Monte Carlo MSE very well with absolute relative biases less than 1.1%.

Turning to the ratio adjusted estimator, we see that the full jackknife variance estimator again tracks the Monte Carlo MSE relatively well for all populations with absolute relative biases less than 5%. The shortcut estimator, on the other hand, has relative biases varying from -67% to 364%. Looking at expression (3.7), we see that for a fixed response rate the bias depends on the $\mathrm{CV}_h(y)$, $\mathrm{CV}_h(x)$ and ρ_{hxy}. Due to the large intercept terms in the second population, $\bar{y}_h$ are large and the corresponding $\mathrm{CV}_h(y)$ are smaller than in the other populations. Thus, the last term in expression (3.7) is quite large and the resulting relative bias of v_{JS} is also large. This is also seen for population 3 except to a lesser extent since only the first stratum has an intercept term. The opposite effect is seen in population 4, where the introduction of zeros has significantly increased $\mathrm{CV}_h(y)$ which has in turn reduced the Monte Carle percent relative bias of the shortcut estimator.

Additional simulations were performed using the some of the populations described in Table 2 but with varying response rates. The results are not presented here as they were as expected. That is, the bias of the shortcut estimator deceased as the response rate increased (with all the other parameters remaining fixed). The full jackknife estimator continued to track the Monte Carlo MSE very well.

5. Conclusion

In this paper, we evaluated both theoretically and empirically a shortcut jackknife variance estimator that does not re-calculate the nonresponse adjustment factors within each jackknife replicate, specifically considering three different nonresponse weighting adjustment procedures. We showed in the context of stratified simple random sampling that the shortcut jackknife variance estimator tends to overestimate the true variance of the estimators. In the context of the ratio adjustment procedure, however, the shortcut jackknife variance estimator may underestimate the true variance if the ratio model is not appropriate for the data at hand.

One justification for the use of a shortcut procedure in a replicate variance estimation method is to save time and computing resources. If these are truly issues and the program has consistently high unit response rates in all weighting cells, then while there are clearly theoretical advantages to replicating the weight adjustment procedure, there may be little or no practical advantage. Having said that, the conditions for "practical" equivalence between the full and shortcut procedure variance estimators are extremely restrictive, and we have demonstrated that small changes in underlying data conditions can easily violate these conditions. If computational concerns with a full jackknife are truly an issue, then the authors recommend the linearization jackknife variance estimation approach which has the same asymptotic properties as the full jackknife, but is computationally quick and computer overhead "free" (in terms of replicate storage). See Thompson and Yung (2006) for expressions for the linearization jackknife variance estimator for both the count and ratio adjusted estimators. Given these viable alternatives, we recommend against the use of a shortcut procedure variance estimator.

Acknowledgements

This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau. The authors would like to thank the Associate Editor, two anonymous referees, Samson Adeshiyan, Patrick Cantwell, Carol Caldwell, Michael Hidiroglou, Rita Petroni, Mark Sands, and Jun Shao for their useful comments on earlier versions of this paper. Work of David Haziza was supported by grants from the Natural Sciences and Engineering Research Council of Canada.

References

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society B*, 67, 445-458.

Brick, M.J., and Montaquila, J.M. (2009). Nonresponse and weighting. In the *Handbook of Statistics*, *Sample Surveys*: *Design*, *Methods and Applications*, (Eds., C.R. Rao and D. Pfeffermann), 29A, 163-185.

Da Silva, D.N., and Opsomer, J.D. (2006). A kernel smoothing method to adjust for unit nonresponse in sample surveys. *Canadian Journal of Statistics*, 34, 563-579.

Deville, J.-C. (2002). La correction de la non-réponse par calage généralisé. *Actes des Journées de Méthodologie Statistique*, Insee.

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

Eltinge, J.L., and Yansaneh, I.S. (1997). Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. Consumer Expenditure Survey. *Survey Methodology*, 23, 33-40.

Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference*, *US Bureau of the Census*, 429-440.

Kalton, G., and Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19, 81-97.

Kim, J.K., and Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*, 35, 501-514.

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

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.

Ozcoskun, L., Thompson, K.J. and Williams, Q. (2005). Investigation of balanced repeated replication (BRR) variance estimation for the Survey of Residential Alterations and Repairs (SORAR). *Proceedings of the Federal Committee on Statistical Methodology*, Office of Management and Budget.

Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: John Wiley & Sons, Inc.

Sautory, O. (2003). CALMAR 2: A new version of the CALMAR calibration adjustment program. *Proceedings*: *Symposium 2003*, *Challenges in Survey Taking for the Next Decade*, Ottawa, Canada.

Shao, J., and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 93, 254-265.

Thompson, K.J. (2005). An empirical investigation into the effects of replicate reweighting on variance estimates for the annual capital expenditures survey. *Proceedings of the Federal Committee on Statistical Methodology*, Office of Management and Budget.

Thompson, K.J., and Yung, W. (2006). To Replicate (A weight adjustment procedure) or not to replicate? An analysis of the variance estimation effects of a shortcut procedure using the stratified jackknife. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 3772-3779.

Valliant, R. (2004). The effect of multiple weighting steps on variance estimation. *Journal of Official Statistics*, 20, 1-18.

Yung, W., and Rao, J.N.K. (2000). Jackknife variance estimation under imputation for estimators using poststratification information. *Journal of the American Statistical Association*, 95, 903-915.

Survey Methodology, June 2010
Vol. 36, No. 1, pp. 45-56
Statistics Canada, Catalogue No. 12-001-X

A comparison of variance estimators for poststratification to estimated control totals

Jill A. Dever and Richard Valliant [1]

Abstract

Calibration techniques, such as poststratification, use auxiliary information to improve the efficiency of survey estimates. The control totals, to which sample weights are poststratified (or calibrated), are assumed to be population values. Often, however, the controls are estimated from other surveys. Many researchers apply traditional poststratification variance estimators to situations where the control totals are estimated, thus assuming that any additional sampling variance associated with these controls is negligible. The goal of the research presented here is to evaluate variance estimators for stratified, multi-stage designs under estimated-control (EC) poststratification using design-unbiased controls. We compare the theoretical and empirical properties of linearization and jackknife variance estimators for a poststratified estimator of a population total. Illustrations are given of the effects on variances from different levels of precision in the estimated controls. Our research suggests (i) traditional variance estimators can seriously underestimate the theoretical variance, and (ii) two EC poststratification variance estimators can mitigate the negative bias.

Key Words: Estimated-control poststratification; Sampling frame coverage bias; Survey-estimated control totals.

1. Introduction

Poststratified estimators, and other calibration estimators, are used in many types of surveys to reduce variances or to correct for frame deficiencies. Specific examples include large U.S. government surveys, such as the Consumer Expenditure Survey (see, *e.g.*, Jayasuriya and Valliant 1996); surveys of specialized populations, such as the U.S. Department of Defense Survey of Health Related Behaviors among Military Personnel (Bray, Hourani, Rae, Dever, Brown, Vincus, Pemberton, Marsden, Faulkner and Vandermaas-Peeler 2003); and a myriad of surveys outside the U.S. including the Canadian Retail Trade Survey (see, *e.g.*, Hidiroglou and Patak 2006), the Swedish Labour Force Survey (Mirza and Hörngren 2002), and the British Household Panel Survey (Taylor, Brice, Buck and Prentice-Lane 2007).

Calibration estimators, such as those generated under poststratification, are used to minimize errors associated with incomplete sampling frames (*i.e.*, undercoverage) and with sampling and nonresponse (see, *e.g.*, Särndal, Swensson and Wretman 1992; Lessler and Kalsbeek 1992; Kott 2006). For example, estimates from the Behavioral Risk Factor Surveillance System (BRFSS), a nationwide random-digit-dial (RDD) telephone survey conducted by the U.S. Centers for Disease Control and Prevention (CDC), are poststratified to counts that include households with and without landline telephone service (Centers for Disease Control and Prevention 2006). The decrease in the errors is linked to the association of the population control totals with the frame undercoverage, patterns of non-ignorable nonresponse, and the variable of interest (Kim, Li and Valliant 2007).

When relevant population controls do not exist, many researchers use survey-estimated control totals, and apply traditional variance formulae as if the controls were known without error. For example, Nadimpalli, Judkins and Chu (2004) adjusted weights for the *2003 National Survey of Parents and Youth* to the number of U.S. households with children ages 9-18 estimated from the *Current Population Survey* (CPS) using a ratio-raking algorithm (www.census.gov/cps). Estimates of how people in the U.S. spend their time can be calculated from *The American Time Use Survey* using weights that have been poststratified to projected estimates from the U.S. decennial Census (Killion 2006). More recently, researchers at the Pew Research Centers calibrated weights for a set of 2008 U.S. presidential pre-election surveys to population estimates from the March 2007 CPS, as well as to estimates on telephone usage patterns from the July-December 2007 *National Health Interview Survey* (Keeter, Dimock and Christian 2008).

The goal of our research is to develop and evaluate variance estimators for point estimates with weights that contain a poststratification adjustment to a set of survey-estimated control totals. We label the methodology which properly accounts for the estimated controls as *estimated-control* (*EC*) *poststratification*. In this paper, we focus specifically on the EC poststratified (ECPS) estimator of a population total for data collected from a stratified, multi-stage design, where the first-stage sampling units are selected *with replacement*. The remainder of this section gives a brief review of weight calibration and poststratification. Section 2

1. Jill A. Dever, RTI International. Email: jdever@rti.org; Richard Valliant, Survey Research Center, University of Michigan and Joint Program in Survey Methodology, University of Maryland. Email: rvalliant@survey.umd.edu.

contains an explicit definition of the ECPS estimator under study, followed in Section 3 by an evaluation of the bias properties. Through a theoretical evaluation (Section 4) and a simulation study, we compare variance estimators developed for the ECPS estimator with a variance estimator chosen under the naïve "population control total" assumption. Both linearization and replication variance estimators are examined in our research. We provide illustrations on the effects of different levels of precision in the estimated controls on the variance estimates. The specifications for the simulation study are detailed in Section 5, followed by a summary of the results (Section 6). We conclude the paper with a brief summary and an overview of future research in this area.

Calibration estimators (Deville and Särndal 1992), such as a poststratified estimator of a population total, borrow strength from auxiliary information to improve the efficiency of survey estimates over simpler weighting methods. When the auxiliary variables are (linearly) related to the set of key survey variables, calibration estimators can be very efficient.

The general form of a *traditional* or *fixed-control* calibration estimator is best described as an expansion estimator or "linear weighting" estimator as discussed in Estevao and Särndal (2000). Define s to be the set of sample elements from a probability sample, and $d_k = 1/\pi_k$ to be the design weight for element k such that $\pi_k = \Pr(k \in s)$. An estimated population total of a variable y is $\hat{t}_y = \sum_{k \in s} w_k y_k$, where the calibration weight $(w_k = a_k d_k)$ for the k^{th} element defined as a function of the design weight, d_k, and a calibration-adjustment factor, a_k, also known as a g-weight (Särndal *et al.* 1992). The calibration weights are calculated by minimizing a specified function that measures the distance between the design and calibration weights subject to a set of constraints defined as:

$$\mathbf{t}_{Ux} = \hat{\mathbf{t}}_{Ax} \qquad (1)$$

where $\mathbf{t}_{Ux} = \sum_{k \in U} \mathbf{x}_k$, the vector of population controls (counts) corresponding to the $G\,(G \geq 1)$ auxiliary variables; $\hat{\mathbf{t}}_x = \sum_{k \in s} w_k \mathbf{x}_k$, the estimated population controls corresponding to the components of $\mathbf{t}_{Ux}$; and $\mathbf{x}_k$ is a vector of length G containing auxiliary or benchmark variable values for element k. Note that $\mathbf{x}_k$ may contain ones and zeros to indicate the presence or absence of a certain characteristic (*e.g.*, age 18-25), or larger values (*e.g.*, number of children). An example of such a calibration system is the generalized least squares (or chi-square) distance function $\sum_{k \in s_A} (w_k - d_k)^2 / c_k d_k$ that is minimized subject to the constraints in (1). This system generates a closed-form solution called the generalized regression estimator (GREG) for $c_k = 1$ (Deville and Särndal 1992). The poststratified estimator is a special case of the GREG.

Variance estimation techniques for the poststratified estimator, and more generally for the GREG, have been widely studied. Binder (1995) demonstrates techniques used to calculate a *Taylor linearization* variance estimator for the GREG. Additional references for the linearization variance estimator under poststratification (and calibration more generally) include Deville, Särndal and Sautory (1993), Demnati and Rao (2004), and Hidiroglou and Patak (2006). Särndal, Swensson and Wretman (1989) developed an approximate linearization variance for the GREG of a population total as a function of the population residuals from a specified model and the design weights (d_k). Valliant (1993) and Yung and Rao (1996) modified the residual-based variance estimator by multiplying the sample residuals by the calibration weights $w_k (= a_k d_k)$. They demonstrated that this revised estimator, created by linearizing the associated jackknife, reduced the bias associated with the original formula. This variance estimator is also discussed in Särndal *et al.* (1992), Stukel, Hidiroglou and Särndal (1996), and in Chapter 11 of Särndal and Lundström (2005). Properties of replication variance estimators (*i.e.*, jackknife and BRR) have been examined in, for example, Valliant (1993), Rust and Rao (1996), Canty and Davison (1999), Théberge (1999), Rao and Shao (1999), Yung and Rao (1996; 2000), and Kott (2006).

An assumption in the articles above is that the control totals, to which the auxiliary sample estimates are adjusted, are either true population values known without error, or are taken from an independent, highly precise survey that is much larger than the survey requiring calibration. In some cases, however, these controls are estimated from other surveys with non-negligible sampling variances. For example, there are efforts to calibrate Web panel surveys to separate, higher-quality reference surveys that are not much larger than the panel surveys themselves (*e.g.*, Krotki 2007; Terhanian, Bremer, Smith and Thomas 2000).

Many researchers apply formulae developed for traditional poststratification even though the controls have been estimated. The tacit assumption is that any additional error (variance and bias) associated with these controls is negligible and can be ignored. Currently, the validity of this assumption can not be checked until a complete picture of EC poststratification has been developed.

2. The estimated-control poststratified estimator

To facilitate our discussion of the estimated-control poststratified estimator, we label the survey requiring poststratification as the *analytic survey* and the source of the control totals as the *benchmark survey*. In practice, more than one benchmark survey may be tapped for the control totals. However, we will assume only one benchmark

survey for the theoretical development so that control total variances and covariances are estimable.

Let U represent the finite target population containing N elements and $t_y = \sum_{k \in U} y_k$ represent the population total of interest for a variable y. Let s_A represent a random sample of size n_A from the frame U_A for the analytic survey. A random sample s_B of size n_B is selected for the benchmark survey from the corresponding sampling frame U_B. We allow the possibility that each of the frames, U_A and U_B, do not completely cover the target population U. However, coverage is treated as a random event so that all elements in the target population have a positive probability of being covered by either the analytic or the benchmark survey frame.

As a convention throughout the paper, an "A" subscript signifies an association with the analytic survey such as a sample design parameter or an estimate. A "B" subscript identifies the benchmark survey quantities. These subscripts are absent from the parameters associated with the population of interest, *i.e.*, t_y.

For the stratified, multi-stage design assumed for the analytic survey, $m_{Ah} (m_{Ah} \geq 2)$ primary sampling units (PSUs), indexed by i, are selected *with replacement* from a total of M_{Ah} PSUs in the h^{th} design stratum ($h = 1, \ldots, H$ with $H \geq 2$). We assume that n_{Ahi} elements, each indexed by k, are selected from N_{Ahi} in PSU hi in such a way that an unbiased estimate of the PSU total can be made. The design weight, d_k, is calculated as the inverse of the unconditional inclusion probability for $k \in s_{Ahi}$, the set of analytic survey elements within the hi^{th} PSU. Thus, n_A, the size of the analytic survey sample, is calculated as $n_A = \sum_{h=1}^{H} \sum_{i=1}^{m_{Ah}} n_{Ahi}$. Elements for the benchmark survey are randomly drawn from the corresponding sampling frame; no explicit specifications are made for the random sampling method.

Poststratification can be used to correct for sampling and coverage errors. Therefore, we allow undercoverage in the analytic-survey, as well as, the benchmark-survey sampling frames. Additionally, we do not consider the effects of nonresponse.

Suppose that the population U can be divided into $g = 1, ..., G$ mutually exclusive and exhaustive poststrata. When the population count of elements, N_g, is known for each poststratum, the traditional poststratified estimator of a total for y is defined as

$$\hat{t}_{yPS} = \sum_{g=1}^{G} N_g \frac{\hat{t}_{Ayg}}{\hat{N}_{Ag}}, \qquad (2)$$

where y_k is the value of the analysis variable y for element k; $\hat{t}_{Ayg} = \sum_{k \in s_A} \delta_{gk} d_k y_k$, the total of y in poststratum g estimated from the analytic survey data; $\hat{N}_{Ag} = \sum_{k \in s_A} \delta_{gk} d_k$, the analytic survey estimated total in poststratum g; and $\delta_{gk} = 1$ indicates membership in the g^{th} poststratum and zero otherwise. Note that $\hat{t}_{Ayg}$ may also be expressed as $\hat{t}_{Ayg} = \sum_{k \in s_{Ag}} d_k y_k$, where s_{Ag} indicates the set of analytic survey elements in poststratum g. The "hat" notation in the expression above is used to distinguish a population estimator (*e.g.*, $\hat{N}_{Ag}$) from the known population parameter (*e.g.*, N_g). If the count of elements in poststratum g is estimated by setting $y_k = 1$ in the formula for $\hat{t}_{Ayg}$, then $\hat{t}_{yPS}$ equals N_g. In this sense, $\hat{t}_{yPS}$ is poststratified to the population counts $N_1, \ldots, N_G$.

In certain situations, however, the population counts are not available and must be estimated from a benchmark survey. Define the ECPS estimator of a population total of a variable y as

$$\hat{t}_{yP} = \sum_{g=1}^{G} \hat{N}_{Bg} \frac{\hat{t}_{Ayg}}{\hat{N}_{Ag}}. \qquad (3)$$

The number of population elements in the g^{th} poststratum $(g = 1, \ldots, G)$ estimated from the benchmark survey is denoted as $\hat{N}_{Bg} = \sum_{l \in s_{Bg}} w_l$, where s_{Bg} is the set of sample elements in poststratum g from the benchmark survey and w_l is the weight associated with the l^{th} element. The calibration-adjustment factors applied to the analytic survey design weights for $\hat{t}_{yP}$ are calculated as $a_k = \hat{N}_{Bg} / \hat{N}_{Ag}$ for $k \in s_{Ag}$.

Relating the poststratified estimators to the calibration system discussed in the previous section, $\hat{\mathbf{t}}_{Ax}$ is a G-length vector of estimated population counts for each poststratum such that $\hat{\mathbf{t}}_{Ax} = (\hat{t}_{Ax1}, \ldots, \hat{t}_{AxG})'$, where $\hat{t}_{Axg} \equiv \hat{N}_{Ag} = \sum_{k \in s_A} d_k \delta_{gk}$ and $x_k \equiv \delta_{gk} = 1$ if the element k is a member of the g^{th} poststratum and 0 otherwise. The vector $\mathbf{t}_{Ux}$ corresponds either to $\mathbf{N} = (N_1, \ldots, N_G)'$ for the $\hat{t}_{yPS}$ estimator given in (2), or to $\hat{\mathbf{N}}_B = (\hat{N}_{B1}, \ldots, \hat{N}_{BG})'$, a $G \times 1$ vector of benchmark control estimates, for the $\hat{t}_{yP}$ estimator given in (3).

The estimator $\hat{t}_{yP}$ can be expressed in matrix notation as $\hat{t}_{yP} = \hat{\mathbf{N}}_B' \hat{\bar{\mathbf{Y}}}_A$ where $\hat{\bar{\mathbf{Y}}}_A = (\hat{\mathbf{N}}_A)^{-1} \hat{\mathbf{t}}_{Ay}$, a $G \times 1$ vector of analytic survey estimates of the form $\hat{\bar{\mathbf{Y}}}_A = [\hat{t}_{A1} / \hat{N}_{A1}, \ldots, \hat{t}_{AG} / \hat{N}_{AG}]'$; $\hat{\mathbf{N}}_A = \text{diag}(\hat{N}_{A1}, \ldots, \hat{N}_{AG})$, a diagonal matrix of poststratum totals estimated from the analytic survey; and $\hat{\mathbf{t}}_{Ay} = [\hat{t}_{A1}, \ldots, \hat{t}_{AG}]'$ is a $G \times 1$ vector of poststratum totals for the outcome variable estimated from the analytic survey. The remaining variables associated with the matrix notation were defined previously.

An effective poststratification adjustment can reduce the bias in the resulting point estimates and will either reduce or minimally inflate the variance in comparison to the unadjusted weight. This effect is well known for traditional poststratification; we provide the comparative evaluation under an estimated-control setting in the next sections.

3. Bias in the ECPS of a population total

Traditional poststratification is known for reducing the bias associated with an incomplete sampling frame. This reduction is most successful when poststrata are formed such that the within-poststratum correlation of y_k with the probability of the k^{th} element being included on the sampling frame is very near zero (Kim, Li and Valliant 2007).

To evaluate the (unconditional) design-based bias for $\hat{t}_{yP}$, we must account for the random property of four components – the analytic and benchmark sample designs and the population coverage propensities for the corresponding sampling frames. Following the work of Kim, Li and Valliant (2007, equation 2), the approximate design bias of $\hat{t}_{yP}$ as an estimator of the population total $t_y = \sum_{k\in U} y_k$ is calculated as

$$\begin{aligned}\text{Bias}(\hat{t}_{yP}) &= E(\hat{t}_{yP}) - t_y \\ &\cong \sum_{g=1}^{G}\left[t_{yg}\left\{\frac{N_{Bg}}{N_g} - 1\right\} + N_{Bg}\text{Cov}(y_g, \phi_{Ag})\overline{\phi}_{Ag}^{-1}\right] \quad (4)\end{aligned}$$

where N_g is the population size for the set of elements U_g within poststratum g; $N_{Bg} = E(\hat{N}_{Bg})$, the expected value of the poststratum estimates under the benchmark survey design; $\text{Cov}(y_g, \phi_{Ag}) = N_g^{-1}\sum_{k\in U_g}(y_k - \overline{y}_g)(\phi_{Ak} - \overline{\phi}_{Ag})$, the population covariance between the outcome variable (y_k) and the coverage propensities (ϕ_{Ak}) within poststratum g; $\overline{y}_g = t_{yg}/N_g$, the g^{th} poststratum mean of y; $t_{yg} = \sum_{k\in U_g} y_k$, the population total of y within poststratum g; and $\overline{\phi}_{Ag} = N_{Ag}/N_g$, the average coverage propensity within the poststratum under the analytic survey design with $N_{Ag} = E(\hat{N}_{Ag})$. Note that the population total may also be expressed as $t_y = \sum_g t_{yg}$.

Components of the bias are zero only under certain conditions. (*i*) If $N_{Bg} = N_g$ for all g (*i.e.*, no coverage errors in the benchmark sampling frame), then the bias is dependent only on the association between the outcome variable and the coverage propensities, $\text{Cov}(y_g, \phi_{Ag})$. The value of $\text{Bias}(\hat{t}_{yP})$ then reduces to the formula provided in Kim, Li and Valliant (2007, equation 2) for the traditional poststratified estimator, $\hat{t}_{yPS}$. (*ii*) If the coverage probabilities are constant within each poststratum (*i.e.*, $\phi_{Ak} = \overline{\phi}_{Ag}$, $k \in U_g$ for all g), then the second bias component is zero. Only if *both* conditions are satisfied can we say that $\hat{t}_{yP}$ is approximately unbiased. Some may argue that a "perfect" combination of poststrata could be formed such that the positive and negative components cancel; however, we believe this likelihood to be so rare as to be virtually impossible.

Having examined bias, we present an evaluation of the variance of $\hat{t}_{yP}$. For some estimators, the contribution of the bias (squared) to the total mean square error (MSE) is small relative to the variance.

4. Variance estimation for the ECPS

Variance estimators have been developed for traditional poststratification and are available in software designed to analyze survey data, *e.g.*, R® (R Development Core Team 2009), SAS® (SAS Institute Inc. 2009), Stata® (StataCorp 2010), and SUDAAN® (Research Triangle Institute 2008). However, limited work has been completed on variance estimation for EC poststratification.

Four EC variance estimators for $\hat{t}_{yP}$ that account for the variance in the control totals are presented in the following subsections after defining the population sampling variance. They include one newly developed linearization variance estimator, and three delete-one-PSU (delete-one) jackknife variance estimators. With the delete-one jackknife, replicates are created by sequentially deleting one PSU and adjusting the weights for the remaining PSUs within the corresponding design stratum. This results in a total of $m_A = \sum_{h=1}^{H} m_{Ah}$ replicates calculated by summing the number of analytic-survey PSUs per stratum (m_{Ah}) across the H strata $(h = 1, ..., H)$.

An effective variance estimator will reproduce the corresponding population sampling variance in expectation. The approximate (or asymptotic) population sampling variance of $\hat{t}_{yP} = \hat{\mathbf{N}}_B'\hat{\overline{\mathbf{Y}}}_A$ has the following form:

$$\begin{aligned}\text{AV}(\hat{t}_{yP}) &= \mathbf{N}_B'\mathbf{V}_A\mathbf{N}_B + 2\overline{\mathbf{Y}}_A'\text{Cov}(\hat{\mathbf{N}}_B, \hat{\overline{\mathbf{Y}}}_A)\mathbf{N}_B + \overline{\mathbf{Y}}_A'\mathbf{V}_B\overline{\mathbf{Y}}_A \\ &= \mathbf{N}_B'\mathbf{V}_A\mathbf{N}_B + \overline{\mathbf{Y}}_A'\mathbf{V}_B\overline{\mathbf{Y}}_A \quad (5)\end{aligned}$$

where $\mathbf{N}_B = E(\hat{\mathbf{N}}_B)$, a vector of expected values for the benchmark poststratum counts within the G poststrata; $\hat{\mathbf{N}}_B = (\hat{N}_{B1}, \ldots, \hat{N}_{BG})'$ is a G-length vector of control totals estimated from the benchmark survey; $\overline{\mathbf{Y}}_A$ is a G-length vector with population components of the form $\overline{y}_{Ag} = t_{Ayg}/N_{Ag}$; $\mathbf{V}_A$ is the population (variance-)covariance matrix of the estimated components of the vector $\overline{\mathbf{Y}}_A$; and $\mathbf{V}_B$ is the covariance matrix of the G benchmark control estimates $\hat{\mathbf{N}}_B$. The first component, $\mathbf{N}_B'\mathbf{V}_A\mathbf{N}_B$, is the approximate variance for the traditional poststratified estimator $\hat{t}_{yPS}$, *i.e.*, the benchmark estimates are treated as fixed. The component, $\overline{\mathbf{Y}}_A'\mathbf{V}_B\overline{\mathbf{Y}}_A$, is the variance associated with the benchmark estimates conditioned on the analytic survey sample; this is the EC poststratification variance component. Because we assume that the analytic and benchmark surveys are independent, the covariance of estimates from the two surveys is, by definition, zero. Hence, the component $\text{Cov}(\hat{\mathbf{N}}_B, \hat{\overline{\mathbf{Y}}}_A)$ above is eliminated from the expression.

Krewski and Rao (1981), Rao and Wu (1985), and others demonstrated the asymptotic consistency of the linearization and jackknife variance estimators for nonlinear functions. However, this examination needs to be extended to the EC poststratification. We discuss the set of EC variance

estimators for the population sampling variance below identified or developed for our research. The sample estimators were calculated by substituting sample estimates for the corresponding variance parameters. We begin with an evaluation of a traditional or naïve poststratified variance estimator that does not account for the variation in the estimated controls.

4.1 A traditional variance estimator for EC poststratification (Naïve)

A variety of variance estimators have been developed for poststratification estimators. With all of the methods, the controls are assumed to be fixed and known without error. Therefore, $\bar{\mathbf{Y}}_A' \mathbf{V}_B \bar{\mathbf{Y}}_A$, the second (positive) component in expression (5), is zero because $\mathbf{V}_B = \mathbf{0}$ by assumption. The linearization variance estimator has the form

$$\text{var}_{\text{Naïve}}(\hat{t}_{yP}) = \hat{\mathbf{N}}_B' \hat{\mathbf{V}}_A \hat{\mathbf{N}}_B \tag{6}$$

where $\hat{\mathbf{N}}_B$ is the vector of the G benchmark control total estimates, and $\hat{\mathbf{V}}_A$ is the estimated covariance matrix of the estimates $\hat{\bar{\mathbf{Y}}}_A = (\hat{t}_{Ay1}/\hat{N}_{A1}, \ldots, \hat{t}_{AyG}/\hat{N}_{AG})$. Because the second component in the second line of (5) is not estimated, any variance formula developed for traditional post-stratification will by definition underestimate the population sampling variance. However, highly precise benchmark estimates may contribute a negligible EC-poststratification variance component to the overall estimate. Thus, the difference between the estimates for traditional and EC poststratification will for these situations also be negligible.

4.2 Taylor series linearization (ECTS)

A linearization variance estimator for the $\hat{t}_{yP}$ has the form:

$$\text{var}_{\text{ECTS}}(\hat{t}_{yP}) = \hat{\mathbf{N}}_B' \hat{\mathbf{V}}_A \hat{\mathbf{N}}_B + \hat{\bar{\mathbf{Y}}}_A' \hat{\mathbf{V}}_B \hat{\bar{\mathbf{Y}}}_A \tag{7}$$

where $\hat{\mathbf{V}}_B$ is the estimated benchmark covariance matrix for the set of G control totals. The remaining terms are defined for expression (6). The ECTS formula is a function of the variance under traditional poststratification and an additive inflation term associated with the variation in the benchmark controls, *i.e.*, $\text{var}_{\text{ECTS}}(\hat{t}_{yP}) = \text{var}_{\text{Naïve}}(\hat{t}_{yP}) + \hat{\bar{\mathbf{Y}}}_A' \hat{\mathbf{V}}_B \hat{\bar{\mathbf{Y}}}_A$.

Ideally, the benchmark survey analysis file would be available to calculate the values for $\hat{\mathbf{V}}_B$. However, researchers may have to rely on published estimates for only the marginal control totals, *i.e.*, point and variance estimates by one characteristic instead of the counts and covariance estimates for a set of characteristics. The implications of having limited information are discussed further in Section 4.4.

4.3 Fuller two-phase jackknife method (ECF2)

Isaki, Tsay and Fuller (2004) applied a two-phase delete-one jackknife variance estimator developed by Fuller (1998) to an EC poststratification situation. The premise behind Fuller's methodology (ECF2) is to take a spectral (eigenvalue) decomposition of the benchmark covariance matrix ($\hat{\mathbf{V}}_B$), develop benchmark adjustments that are a function of the resulting eigenvalues and eigenvectors, and add the adjustments to the vector of benchmark controls ($\hat{\mathbf{N}}_B$) to create a set of replicate controls. A randomly chosen subset of the m_A replicates is poststratified to the G constructed replicate controls where the total number of PSUs must equal or exceed the number of poststrata, *i.e.*, $m_A \geq G$. Specifically, the benchmark control total for the r^{th} replicate is defined as

$$\hat{\mathbf{N}}_{B(r)} = \hat{\mathbf{N}}_B + c_h \hat{\mathbf{z}}_{(r)}' \tag{8}$$

where $\hat{\mathbf{z}}_{(r)}' = \delta_{(r)} \sum_{g=1}^{G} \delta_{g|(r)} \hat{\mathbf{z}}_g'$; $c_h = \sqrt{m_{Ah}/(m_{Ah}-1)}$, a constant related to the delete-one jackknife variance method; $\delta_{(r)}$ is a zero/one indicator that identifies the G (out of m_A) randomly chosen replicates to receive an adjustment; $\delta_{g|(r)} = 1$ if the g^{th} component of the benchmark covariance decomposition is randomly chosen for the assignment given that replicate r is selected for adjustment; and $\hat{\mathbf{z}}_g = \hat{\mathbf{q}}_g \sqrt{\hat{\lambda}_g}$, a function of an eigenvector ($\hat{\mathbf{q}}_g$) and the associated eigenvalue ($\hat{\lambda}_g$) where $\hat{\mathbf{V}}_B = \sum_{g=1}^{G} \hat{\mathbf{z}}_g \hat{\mathbf{z}}_g'$, by definition. Thus, given that $\delta_{(r)} = 1$ for a particular replicate, a single indicator $\delta_{g|(r)}$ must also equal one; however, if $\delta_{(r)} = 0$, then *all* indicators $\delta_{g|(r)}$ equal zero.

The delete-one jackknife can take multiple forms depending on the centering value. We chose the somewhat conservative variance estimator centered about the full-sample estimate for our research (v_4 in Wolter 2007, section 4.5). The delete-one jackknife variance estimator, $\text{var}_{\text{ECF2}}(\hat{t}_{yP})$, is calculated as follows under the Fuller method for a stratified, multi-stage design.

$$\begin{aligned}\text{var}_{\text{ECF2}}(\hat{t}_{yP}) &= \sum_{h=1}^{H} \frac{(m_{Ah}-1)}{m_{Ah}} \sum_{r=1}^{m_{Ah}} (\ddot{t}_{yP(r)} - \hat{t}_{yP})^2 \\ &= \sum_{h=1}^{H} \frac{(m_{Ah}-1)}{m_{Ah}} \sum_{r=1}^{m_{Ah}} (\hat{t}_{yP(r)} - \hat{t}_{yP} + c_h \hat{\mathbf{z}}_{(r)}' \hat{\mathbf{B}}_{A(r)})^2 \end{aligned} \tag{9}$$

where the terms in (9) are defined below. Note that the association of the r^{th} replicate to a particular design stratum is defined through the stratum membership of the eliminated PSU. The replicate estimates in (9) are defined as $\hat{t}_{Ayg(r)} = \sum_h \sum_{i \in s_{Ah}} d_{i(r)} \sum_{k \in s_{Ahi}} \delta_{gk} d_k y_k$ and $\hat{N}_{Ag(r)} = \sum_h \sum_{i \in s_{Ah}} d_{i(r)} \sum_{k \in s_{Ahi}} \delta_{gk} d_k$, where the PSU-subsampling weights are calculated as

$$d_{i(r)} = \begin{cases} 0 & \text{if } r=i,\ i\in s_{Ah} \\ 1 & \text{if } h\neq h' \text{ for } r\in s_{Ah} \text{ and } i\in s_{Ah'} \\ m_{Ah}/(m_{Ah}-1) & \text{if } r\neq i \text{ but } h=h'. \end{cases} \quad (10)$$

The remaining terms in (9) are $\hat{\mathbf{B}}_{A(r)} = \hat{t}_{Ayg(r)}/\hat{N}_{Ag(r)}$, the estimated mean of the outcome variable within poststratum g and replicate r;

$$\ddot{t}_{yP(r)} = \sum_{g=1}^{G} \hat{N}_{Bg(r)} (\hat{t}_{Ayg(r)}/\hat{N}_{Ag(r)}), \quad (11)$$

a function of replicate estimates with $\hat{N}_{Bg(r)}$ defined as the g^{th} component in expression (8); $\hat{t}_{yP(r)}$ is the replicate estimate under traditional poststratification, namely $\sum_{g=1}^{G} \hat{N}_{Bg}(\hat{t}_{Ayg(r)}/\hat{N}_{Ag(r)})$; and $\hat{t}_{yP}$ is the estimated total given in expression (3) calculated from the complete sample file. Squaring the terms in (9) results in a variance component conditioned on the benchmark controls, a component due to the benchmark control variability, and a cross-term of lower order that is approximately equal to zero in expectation. The design-expectation of the resulting jackknife variance estimator is asymptotically equivalent to $\text{AV}(\hat{t}_{yP})$ in (5) only if the respective components are calculated with values from design-consistent estimators. Fuller (1998) also demonstrated that the jackknife variance of the replicate controls, $\text{var}_{\text{ECF2}}(\hat{\mathbf{N}}_B)$, reproduces the estimated benchmark covariance matrix $\hat{\mathbf{V}}_B$ for every sample.

Currently no software exists to calculate the ECF2. The six steps needed to calculate $\text{var}_{\text{ECF2}}(\hat{t}_{yP})$ using any appropriate programmable package are as follows:

1. Calculate the full-sample estimate $\hat{t}_{yP}$ using expression (3).
2. Determine the G eigenvalues $\hat{\lambda}_g$ and eigenvectors $\hat{\mathbf{q}}_g$ for $\hat{\mathbf{V}}_B$, and calculate the replicate adjustments $\hat{\mathbf{z}}_g = \hat{\mathbf{q}}_g\sqrt{\hat{\lambda}_g}$. Concatenate the $G\times G$ matrix of $\hat{\mathbf{z}}_g$'s with a $G\times(m_A - G)$ matrix of zeros, and randomly sort the columns. Call this new $G\times m_A$ matrix $\hat{\mathbf{Z}}$.
3. Calculate a vector of length m_A with values equal to $c_h = \sqrt{m_{Ah}/(m_{Ah}-1)}$ ordering from $h=1$ to H. Populate each row of a $G\times m_A$ matrix, called $\mathbf{C}$, with this vector, *i.e.*, the row values are repeated . The m_A-length vector of jackknife stratum weights, $\mathbf{W}_R$, is created with components equal to $(m_{Ah}-1)/m_{Ah}$ where the deleted PSU is extracted from stratum h.
4. Calculate the Hadamard (or element-wise) product (Searle 1982, page 49) of $\hat{\mathbf{Z}}$ and $\mathbf{C}$ denoted as $\hat{\mathbf{Z}}\bullet\mathbf{C}$. Replicate the vector $\hat{\mathbf{N}}_B$ into the columns of a $G\times m_A$ matrix and add to $\hat{\mathbf{Z}}\bullet\mathbf{C}$. This new $G\times m_A$ matrix, called $\hat{\mathbf{N}}_{BR}$, contains the replicate benchmark controls discussed in expression (8) for all m_A replicates.
5. Calculate the replicate estimates $\hat{\bar{y}}_{Ag(r)} = \hat{t}_{Ayg(r)}/\hat{N}_{Ag(r)}$ by removing in-turn one PSU from the analytic survey sample file, adjusting the weights for the remaining PSUs ($\mathbf{W}_R$ values), and summing the weighted values for the numerator and denominator within poststratum g. Call the resulting $G\times m_A$ matrix $\hat{\bar{\mathbf{Y}}}_R$.
6. Calculate the m_A replicate estimates, $\ddot{t}_{yP(r)}$, by first multiplying the elements $\hat{\mathbf{N}}_{BR}$ by $\hat{\bar{\mathbf{Y}}}_R$ and summing down the rows within a column. Next, subtract $\hat{t}_{yP}$ from each of the m_A values and square the terms, multiply by the PSU-subsampling weight adjustments specified in (10), and sum across the m_A estimates. The resulting value is the estimated variance using the Fuller method, $\text{var}_{\text{ECF2}}(\hat{t}_{yP})$.

4.4 Nadimpalli-Judkins-Chu jackknife method (ECNJC)

Nadimpalli *et al.* (2004) developed a delete-one jackknife variance estimator that randomly perturbs the control totals for the complete set of replicates instead of adjusting only a subsample of replicates as discussed for the ECF2. The benchmark survey replicate control totals have the following form:

$$\hat{\mathbf{N}}_{B(r)} = \hat{\mathbf{N}}_B + c_h R_h \hat{\mathbf{S}}_B \boldsymbol{\eta}_{(r)} \quad (12)$$

where $c_h = \sqrt{m_{Ah}/(m_{Ah}-1)}$, as with the ECF2; $R_h = \sqrt{1/(H\,m_{Ah})}$, a function of the total number of analytic-survey strata (H) and PSUs (m_{Ah}); $\hat{\mathbf{S}}_B$ is a *diagonal* matrix of estimated standard errors for the benchmark controls; and $\boldsymbol{\eta}_{(r)}$ is a G-length vector of values randomly generated for each replicate from the standard normal distribution. The remaining terms are specified for the ECF2 following expression (8). Note that the covariance estimates included in the ECF2, *i.e.*, the off-diagonal values of $\hat{\mathbf{V}}_B$, are set to zero for the ECNJC.

The corresponding delete-one jackknife variance estimator of the poststratified total is calculated as follows:

$$\begin{aligned} \text{var}_{\text{ECNJC}}(\hat{t}_{yP}) &= \sum_{h=1}^{H} \frac{(m_{Ah}-1)}{m_{Ah}} \sum_{r=1}^{m_{Ah}} (\ddot{t}_{yP(r)} - \hat{t}_{yP})^2 \\ &= \sum_{h=1}^{H} \frac{(m_{Ah}-1)}{m_{Ah}} \sum_{r=1}^{m_{Ah}} (\hat{t}_{yP(r)} - \hat{t}_{yP} \\ &\qquad + c_h R_h \boldsymbol{\eta}'_{(r)} \hat{\mathbf{S}}_B \hat{\mathbf{B}}_{A(r)})^2, \end{aligned} \quad (13)$$

where $\ddot{t}_{yP(r)}$ is computed as described for the ECF2 in (11) but with $\hat{N}_{Bg(r)}$ defined by the g^{th} component in (12). Unlike the ECF2, the sample variance of the ECNJC

replicate controls given in (12) reproduces the benchmark covariance matrix $\mathbf{V}_B$ *in expectation* only if the covariance terms are truly zero (see Appendix A for details). If $\mathbf{V}_B$ is not diagonal, $\text{var}_{\text{ECNJC}}$ fails this test.

Use of the ECNJC would be plausible in two cases: (*i*) the complete benchmark covariance matrix for the controls is unavailable (*e.g.*, estimates taken from a previous report), or (*ii*) the covariance terms are negative so that the resulting values defined by (12) would lead to conservative variance estimates. The diagonal matrix for $\hat{\mathbf{S}}_B$ would be correct if the estimated poststratum counts were actually uncorrelated. However this is unlikely because of the multinomial structure of $\hat{\mathbf{N}}_B$. Given the setup for the ECNJC, the expectation of the variance estimator will *not* approximate $\text{AV}(\hat{t}_{yP})$ in (5); the bias term is related to the difference between the design expectation of $\hat{\mathbf{S}}_B^2$ and $\mathbf{V}_B$.

4.5 Multivariate normal jackknife method (ECMV)

The multivariate normal method (ECMV) is a generalization of the ECNJC and to our knowledge is first discussed in this paper. The ECMV uses the complete covariance matrix $\hat{\mathbf{V}}_B$ and relies on large-sample theory so that the control total adjustments may be modeled as coming from a *G*-dimensional multivariate normal (MVN) distribution. The replicate controls for the ECMV have the form

$$\hat{\mathbf{N}}_{B(r)} = \hat{\mathbf{N}}_B + c_h R_h \hat{\boldsymbol{\varepsilon}}_{(r)} \qquad (14)$$

where $\hat{\boldsymbol{\varepsilon}}_{(r)} \underset{\text{i.i.d.}}{\sim} \text{MVN}_G(\mathbf{0}, \hat{\mathbf{V}}_B)$; $c_h = \sqrt{m_{Ah}/(m_{Ah}-1)}$; and $R_h = \sqrt{1/(H\, m_{Ah})}$.

The delete-one jackknife variance estimator for the ECMV is calculated as

$$\begin{aligned}\text{var}_{\text{ECMV}}(\hat{t}_{yP}) &= \sum_{h=1}^{H} \frac{(m_{Ah}-1)}{m_{Ah}} \sum_{r=1}^{m_{Ah}} (\ddot{t}_{yP(r)} - \hat{t}_{yP})^2 \\ &= \sum_{h=1}^{H} \frac{(m_{Ah}-1)}{m_{Ah}} \sum_{r=1}^{m_{Ah}} (\hat{t}_{yP(r)} - \hat{t}_{yP} \\ &\qquad + c_h R_h \hat{\boldsymbol{\varepsilon}}'_{(r)} \hat{\mathbf{B}}_{A(r)})^2, \end{aligned} \qquad (15)$$

where $\ddot{t}_{yP(r)}$ is computed as described for the ECF2 in (11) but with $\hat{N}_{Bg(r)}$ defined by the g^{th} component in (14). Unlike the Fuller method, $\text{var}_{\text{ECMV}}(\hat{\mathbf{N}}_B) \neq \hat{\mathbf{V}}_B$; instead, the ECMV must rely on the design-based properties of the estimator. The design expectation of this estimator is evaluated with respect to the MVN distribution conditioned on the benchmark estimates (E_ε), and then with respect to the benchmark survey design (E_B). As shown in Appendix B.1,

$$E_B[E_\varepsilon(\text{var}_{\text{ECMV}}(\hat{\mathbf{N}}_B)|B)] = E_B(\hat{\mathbf{V}}_B). \qquad (16)$$

If $\hat{\mathbf{V}}_B$ is an approximately unbiased estimator of $\mathbf{V}_B$, then the population covariance matrix is reproduced with this method.

Under the Fuller two-phase method, $\text{Var}[\text{var}_{\text{ECF2}}(\hat{\mathbf{N}}_B)] = \text{Var}(\hat{\mathbf{V}}_B)$ because $\text{var}_{\text{ECF2}}(\hat{\mathbf{N}}_B) = \hat{\mathbf{V}}_B$. To compare ECF2 and ECMV further, note that if we define $y_k = 1$ in the analytic survey, then $\hat{t}_{yP} = \mathbf{1}'\hat{\mathbf{N}}_B$. As shown in Appendix B.2,

$$\text{Var}[\text{var}_{\text{ECMV}}(\mathbf{1}'\hat{\mathbf{N}}_B)] =$$

$$\text{Var}_B[\mathbf{1}'\hat{\mathbf{V}}_B\mathbf{1}] + \frac{2}{H\bar{m}_A^*}[E_B(\mathbf{1}'\hat{\mathbf{V}}_B\mathbf{1})^2] > \text{Var}_B[\mathbf{1}'\hat{\mathbf{V}}_B\mathbf{1}] \qquad (17)$$

where $\bar{m}_A^*$ is the harmonic mean of the PSU sample sizes per stratum in the analytic survey. This suggests that the var_{ECF2} and the var_{ECMV} have similar large sample expectations, though in practice the ECMV is likely to be more variable than the ECF2. We examine this issue through a simulation study described in the next section.

5. Description of simulation study

We complement the theoretical evaluation of the five variance estimators discussed in the previous section with an analysis of simulation results.

5.1 Simulation parameters

The simulation population is a random subset of the 2003 National Health Interview Survey (NHIS) public-use file containing records for 21,664 adults. These records were divided into 25 strata, each containing six PSUs. Samples were selected from this "population" using a two-stage design. Two PSUs were selected *with replacement* using probabilities proportional to the total number of adults (PPS) within the PSU. From within each sample PSU, we selected simple random samples of $(n_{Ahi} =)$ 20 and 40 persons *without replacement* giving total sample sizes of 1,000 and 2,000, respectively. Two within-PSU sample sizes were considered for this study to evaluate the effects of smaller analytic survey variance components, calculated by increasing n_A, on the variance of $\hat{t}_{yP}$. For each combination of PSU and person-level samples (*i.e.*, 50 PSUs and either 1,000 or 2,000 persons), we selected 4,000 simulation samples. We calculated the estimated population totals and associated variances for two binary NHIS variables: NOTCOV = 1 indicates that an adult *did not* have health insurance coverage in the 12 months prior to the NHIS interview (approximately 17 percent of the population); and PDMED12M = 1 indicates that an adult *delayed* medical care because of cost in the 12 months prior to the interview (approximately 7 percent of the population).

We exclude nonresponse from consideration in our current simulation study to minimize factors that might affect our comparisons. (Note: The interview questions for these variables can be found in the family core instrument at ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Survey_Questionnaires/NHIS/2003/qfamilyx.pdf. Responses from questions FHI.070 and FAU.010/FAU.020 were used to generate the variables NOTCOV and PDMED12M, respectively).

Poststratification may reduce variances slightly. However, in household surveys, this technique is mainly used to correct for sampling frame undercoverage, as well as other problems inherent with surveys. Each of the 4,000 simulation samples was selected to mimic a sampling frame for the analytic survey that suffers from differential undercoverage, such as those used for many telephone surveys. Sixteen ($G = 16$) poststratification cells were defined by an eight-level age variable crossed with gender. The coverage rates for the 16 cells were created based on the population means for each age group by gender and range in value from 0.5 to 0.9. A coverage rate equal to 1.0 would indicate full coverage. Before each sample was selected, the frame was designated as a stratified random subsample of the full population of 21,664. For example, 90 percent of the male population 65-69 years of age was randomly selected to be in the sampling frame for the NOTCOV simulations. This process of subsetting the population to the frame was independently implemented for each sample and for each outcome variable.

We suspect that the decision for researchers to use either a traditional or an EC poststratification variance estimator depends on the precision of the control totals. We calculated the benchmark covariance matrix ($\hat{\mathbf{V}}_B$) from the complete NHIS public-use data file (92,148 records) and ratio adjusted the values to reflect a sample size comparable with our simulation population ($N = 21{,}664$). The off-diagonal values of $\hat{\mathbf{V}}_B$ range from -0.05 to 0.75 with a mean value of 0.22. From this matrix we calculated four covariance matrices for the simulation by dividing the original matrix by the adjustment factors 1.0, 3.6, 18, and 72. The adjustments reflect benchmark surveys with an approximate effective sample size of 21,700, 6,000 (≈ 21,700/3.6), 1,200, and less than 500, respectively.

The simulation was conducted in R® (Lumley 2009; R Development Core Team 2009) because of its extensive capabilities for analyzing survey data and efficiency with simulated analyses. Code was developed to calculate the linearization and replicate variance estimates for the EC poststratified estimator discussed above because the relevant code does not currently exist.

5.2 Evaluation criteria

The empirical results for the five variance estimators discussed in the previous section (Naïve, ECTS, ECF2, ECNJC, and ECMV) are compared using three measures across the $j = 1, \ldots, 4{,}000$ simulation samples, and the two outcome variables (NOTCOV and PDMED12M). The measures include: (*i*) the estimated percent relative bias of the variance estimator, $(1/4{,}000\sum_j \mathrm{var}(\hat{t}_{yP_j}) - \mathrm{mse})/\mathrm{mse}$ where $\mathrm{var}(\hat{t}_{yP_j})$ is one of the five variance estimates evaluated for sample *j* and *mse* is the mean square error of $\hat{t}_{yP}$ defined below; (*ii*) the 95% confidence interval coverage rate, $1/4{,}000\sum_j I(|\hat{z}_j| \le z_{1-\alpha/2})$ where $\hat{z}_j = (\hat{t}_{yP_j} - t_y)/\sqrt{\mathrm{var}(\hat{t}_{yP_j})}$; and, (*iii*) the standard deviation of the estimated standard errors, calculated as the square root of $1/(4{,}000-1)\sum_j(\sqrt{\mathrm{var}(\hat{t}_{yP_j})} - 1/4{,}000\sum_j\sqrt{\mathrm{var}(\hat{t}_{yP_j})})^2$. The relative bias and the root mean square error of our point estimators are calculated as $1/4{,}000\sum_j(\hat{t}_{yP_j} - t_y)/t_y$ and $\sqrt{\mathrm{mse}} = \sqrt{1/4{,}000\sum_s(\hat{t}_{yP_j} - t_y)^2}$, respectively.

6. Simulation study results

6.1 Point estimator

To justify the need for poststratification, we initially evaluated the Horvitz-Thompson estimate ($\sum_{s_A} d_k y_k$) for the two outcome variables. This estimator is known to be design-unbiased under pristine conditions. The percent relative bias indicates that the HT estimator is negatively biased, underestimating the population total by 38 percent for NOTCOV and 41 percent for PDMED12M. These large values show that some correction is needed to adjust for the non-negligible levels of bias. The percent relative bias for the poststratified estimator $\hat{t}_{yP}$ was much lower – the $\hat{t}_{yP}$ is positively biased by no more than two percent for both outcome variables.

6.2 Variance estimators

Adding to the theoretical evaluation discussed in Section 4, the empirical results for an effective variance estimator should possess a *percent relative bias* either near zero or somewhat positive for a conservative measure (see Section 5.2 for the formula of the percent relative bias).

The percent relative biases generated from our simulation study are provided in Table 1. Bias estimates for the Naïve and ECNJC variance estimators are larger than for the other EC estimators for all our simulations. Estimates for the ECTS are somewhat smaller than the values calculated for the ECF2 and ECMV estimators for relatively small benchmark surveys. However, the differences are negligible as the size of the benchmark survey increases.

Table 1
Percent relative bias estimates for five variance estimators by outcome variable and relative size of the benchmark survey to the analytic survey

Outcome Variable	Variance Estimator	Relative Size (n_A = 1,000)				Relative Size (n_A = 2,000)			
		0.3	1.2	6.0	21.7	0.2	0.6	3.0	10.8
NOTCOV	Naïve	-50.3	-23	-10.7	-9.2	-56.0	-31	-14.2	-12.2
	ECTS	-4.5	-4.5	-6.1	-7.7	-0.2	-8.4	-8.2	-10.1
	ECF2	-4.7	-4.6	-5.8	-7.5	0.1	-8.2	-8.3	-10.1
	ECNJC	-36.7	-17.1	-8.9	-8.2	-40	-24.2	-11.9	-11.1
	ECMV	-4.3	-4.1	-6.0	-7.5	-0.2	-8.1	-8.1	-10.0
PDMED12M	Naïve	-34.4	-14.5	-5.7	-3.9	-48.1	-23.4	-10	-10.1
	ECTS	-3.3	-3.7	-2.7	-2.6	-4.7	-6.4	-5.1	-7.8
	ECF2	-3.5	-3.5	-2.4	-2.3	-4.6	-6.8	-5.2	-7.8
	ECNJC	-24.5	-10.5	-4.0	-2.7	-35.1	-17.6	-7.6	-8.4
	ECMV	-3.0	-3.3	-2.4	-2.2	-4.3	-6.3	-5.0	-7.7

The traditional poststratified estimator (Naïve) was most negatively biased among those compared as expected. When the benchmark survey is smaller than the analytic survey (and therefore produces estimates less precise than the analytic survey), the Naïve estimator is negatively biased by as much as 56 percent. The level of bias improved as the relative size of the benchmark survey increased; however, the Naïve estimator still resulted in, at best, a four percent underestimate. The ECNJC estimator fared slightly better than the Naïve estimator though the bias (-2.7 to -40 percent) is still larger than the other EC variance estimators, which range between -10.1 and 0.1 percent.

For a small benchmark survey relative to the size of the analytic survey (*i.e.*, relative size less than one), the levels of (absolute) bias dramatically increased for the Naïve and ECNJC estimators. The opposite effect is noted for the other EC variance estimators. The variance component associated with the benchmark survey, *e.g.*, $\hat{\mathbf{Y}}_A' \hat{\mathbf{V}}_B \hat{\mathbf{Y}}_A$ shown for var_{ECTS} in (7), becomes the dominate term within the EC variance estimators as the precision of the benchmark survey estimates decreases. Thus the benchmark variance component somewhat corrects for the underestimation associated with the analytic variance component. Additional research is needed to determine if a threshold exists for when such a counterbalance of bias can occur. The overall negative bias of our estimates is similar to the bias of linearization variance estimators as shown in another context by Rao and Wu (1985, section 4) and Wu (1985). However, further research is also needed to determine how to minimize the underestimation.

Note that the relative sizes of 21.7 when $n_A = 1{,}000$ and 10.8 when $n_A = 2{,}000$ both imply benchmark survey sample sizes of about 21,600. Thus the $O(M^2/m_B)$ component of the variance, $\bar{\mathbf{Y}}_A' \mathbf{V}_B \bar{\mathbf{Y}}_A$, is more prominent for the estimates in Table 1 based on $n_A = 2{,}000$. This leads to larger relative biases in these estimates, relative to those produced under $n_A = 1{,}000$, even though the analytic survey sample size is larger.

The patterns exhibited for the percent relative bias are reflected in the *coverage rates for the 95 percent confidence intervals* for the estimated totals but are not provided for sake of brevity. The Naïve and ECNJC estimators are more likely to experience confidence intervals coverage rates below 95 percent. These rates approach the appropriate level as the precision of the benchmark survey estimates improves. However, the remaining EC variance estimators had coverage rates near acceptable levels regardless of the relative size of the surveys and therefore are more robust.

The discussion so far suggests that there are minimal theoretical, as well as empirical, differences between the ECTS, ECF2, and ECMV methods. We finally look to the *standard deviation of the estimated standard errors* (SEs) in an attempt to distinguish the estimators. An examination of this variability can provide insight on the (empirical) stability of the variance estimators, *i.e.*, an unstable variance estimator could generate a poor variance estimate based on the nuances of a particular sample. Table 2 contains the percent relative increase in the standard deviations for the ECF2 and the ECMV both in comparison to the ECTS.

The variation in the ECMV variance estimates was noticeably larger than for ECF2 but only for relatively small benchmark surveys. The difference increased as the size of the analytic survey increased. This suggests that the ECF2 may be preferred over the ECMV due to increased stability in the variance estimates. However, further research is being conducted on the threshold for when the instability can affect the estimates.

7. Conclusions and future work

The theoretical and analytical work discussed in this paper support the need for a new methodology to address post-stratification using estimated control totals, *i.e.*, estimated-control (EC) poststratification. Traditional variance estimators can severely underestimate the population sampling variance resulting in, for example, incorrect decisions for hypothesis tests and sub-optimal sample allocations when the design is implemented in the future.

Table 2
Percent increase in instability of variance estimates relative to the ects by outcome variable and relative size of the benchmark survey

	Variance	Relative Size (n_A = 1,000)				Relative Size (n_A = 2,000)			
Outcome Variable	Estimator	0.3	1.2	6.0	21.7	0.2	0.6	3.0	10.8
NOTCOV	ECF2	12.0	5.5	2.3	0.2	15.1	8.4	2.1	0.6
	ECMV	21.2	7.4	1.8	0.3	30.8	8.5	2.4	0.7
PDMED12M	ECF2	7.7	3.8	1.1	0.4	12.0	6.3	2.1	0.7
	ECMV	11.5	4.0	0.9	0.5	22.6	7.6	2.2	1.1

The EC linearization variance estimator var_{ECTS} in expression (7) shows promise for EC poststratification. This estimator is especially effective at reducing the percent relative bias experienced with the Naïve variance estimator in (6) when the benchmark survey is small relative to the analytic survey. The replication variance estimator var_{ECF2} given in (9) is recommended specifically for studies requiring replicate weights such as when public-use analysis files are released without sampling design information to further protect data confidentiality and respondent privacy. The alternative replication estimator var_{ECMV} also performed well and is somewhat easier to implement than var_{ECF2}.

Implementation of the recommended variance estimators requires specialized computer programs because the capabilities are currently not available in standard software. The linearization estimator may be more approachable because implementation involves a modification to available variance estimates, *e.g.*, $\text{var}_{\text{ECTS}}(\hat{t}_{y\text{ECPS}}) = \text{var}_{\text{Naïve}}(\hat{t}_{y\text{ECPS}}) + \hat{\mathbf{Y}}_A' \hat{\mathbf{V}}_B \hat{\mathbf{Y}}_A$. We provide a step-by-step discussion of the procedures required for the var_{ECF2} (see Section 4.3) to facilitate the creation of the computer program.

Extensions to this research to be presented at a later date include a generalization to linear calibration, to other statistics including a ratio-estimated mean, and to domain estimation. We additionally are investigating whether threshold values are identifiable which determine (*i*) when there are negligible differences between traditional and EC variance estimation, and (*ii*) when the benchmark controls are too imprecise to use for calibration. We also plan to investigate the theoretical implications of measurement errors in the analytic as well as the benchmark surveys.

Acknowledgements

This work was competed as part of the first author's doctoral dissertation at the Joint Program in Survey Methodology, University of Maryland. She thanks the members of her committee, Richard Valliant, Phillip Kott, Frauke Kreuter, Stephen Miller and Paul Smith for their guidance. The authors also thank the associate editor and referees for their constructive comments which clarified the presentation.

Appendix A

Derivation of $\text{var}_{\text{ECNJC}}(\hat{\mathbf{N}}_B)$

For the following derivations, let E_ε represent the expectation with respect to a standard normal distribution. All other terms are defined in the body of the paper.

$$\text{var}_{\text{ECNJC}}(\hat{\mathbf{N}}_B) = \sum_{h=1}^{H} \frac{m_{Ah}-1}{m_{Ah}} \sum_{r=1}^{m_{Ah}} (\hat{\mathbf{N}}_{B(r)} - \hat{\mathbf{N}}_B)(\hat{\mathbf{N}}_{B(r)} - \hat{\mathbf{N}}_B)'$$

$$= \frac{1}{H} \hat{\mathbf{S}}_B \left(\sum_{h=1}^{H} \frac{1}{m_{Ah}} \sum_{r=1}^{m_{Ah}} \boldsymbol{K}_{(r)} \right) \hat{\mathbf{S}}_B$$

where $\boldsymbol{K}_{(r)} = \boldsymbol{\eta}_{(r)} \boldsymbol{\eta}_{(r)}'$, a $G \times G$ cross-product matrix of standard normal values; and $\hat{\mathbf{S}}_B^2 = \text{diag}(\hat{\mathbf{V}}_B)$. Because $E_\varepsilon(\boldsymbol{K}_{(r)}) = \mathbf{I}_G$, a G-dimension identity matrix, we have $E_\varepsilon[\text{var}_{\text{ECNJC}}(\hat{\mathbf{N}}_B)] = \text{diag}(\hat{\mathbf{V}}_B)$. Therefore, $\text{var}_{\text{ECNJC}}(\hat{\mathbf{N}}_B)$ does not reproduce $\hat{\mathbf{V}}_B$ in expectation.

Appendix B

Evaluation of the ECMV

For the following derivations, let E_B and Var_B represent the expectation and variance with respect to the benchmark survey sampling design. Also, let E_ε and Var_ε represent the expectation and variance with respect to the G-dimensional multivariate normal distribution, $\text{MVN}_G(\mathbf{0}, \hat{\mathbf{V}}_B)$. All other terms are defined in the body of the paper.

B.1: Derivation of $E[\text{var}_{\text{ECMV}}(\hat{\mathbf{N}}_B)]$ given in (15)

Using expression (14) and $c_h^2 = m_{Ah}/(m_{Ah}-1)$,

$$E[\text{var}_{\text{ECMV}}(\hat{\mathbf{N}}_B)] = E_B \left[E_\varepsilon \left(\sum_{h=1}^{H} \frac{(m_{Ah}-1)}{m_{Ah}} \sum_{r=1}^{m_{Ah}} (\hat{\mathbf{N}}_{B(r)} - \hat{\mathbf{N}}_B)(\hat{\mathbf{N}}_{B(r)} - \hat{\mathbf{N}}_B)' \,\middle|\, B \right) \right],$$

$$= \frac{1}{H} E_B \left[\sum_{h=1}^{H} \frac{1}{m_{Ah}} \sum_{r=1}^{m_{Ah}} E_\varepsilon(\hat{\boldsymbol{\varepsilon}}_{(r)} \hat{\boldsymbol{\varepsilon}}_{(r)}' \,|\, B) \right]$$

$$= \frac{1}{H} \sum_{h=1}^{H} \frac{1}{m_{Ah}} \sum_{r=1}^{m_{Ah}} E_B(\hat{\mathbf{V}}_B) = E_B(\hat{\mathbf{V}}_B).$$

B.2: Derivation of $\mathrm{Var}[\mathrm{var}_{\mathrm{ECMV}}(\hat{\mathbf{N}}_B)]$ given in (15)

When $y_k = 1$ so that $\hat{t}_{yP} = \mathbf{1}'\hat{\mathbf{N}}_B$, $\mathrm{var}_{\mathrm{ECMV}}(\mathbf{1}'\hat{\mathbf{N}}_B) = H^{-1}\sum_{h=1}^{H} m_{Ah}^{-1}\sum_{r=1}^{m_{Ah}} \mathbf{1}'\hat{\boldsymbol{\varepsilon}}_{(r)}\hat{\boldsymbol{\varepsilon}}'_{(r)}\mathbf{1}$. Using the formula for the variance of a quadratic form (Searle 1982, section 13.5), we have

$$
\begin{aligned}
&\mathrm{Var}[\mathrm{var}_{\mathrm{ECMV}}(\mathbf{1}'\hat{\mathbf{N}}_B)] \\
&= \mathrm{Var}_B\left[\frac{1}{H}\sum_{h=1}^{H}\frac{1}{m_{Ah}}\sum_{r=1}^{m_{Ah}} E_\varepsilon(\mathbf{1}'\hat{\boldsymbol{\varepsilon}}_{(r)}\hat{\boldsymbol{\varepsilon}}'_{(r)}\mathbf{1}|B)\right] \\
&\quad + E_B\left[\frac{1}{H^2}\sum_{h=1}^{H}\frac{1}{m_{Ah}^2}\sum_{r=1}^{m_{Ah}} \mathrm{Var}_\varepsilon(\mathbf{1}'\hat{\boldsymbol{\varepsilon}}_{(r)}\hat{\boldsymbol{\varepsilon}}'_{(r)}\mathbf{1}|B)\right] \\
&= \mathrm{Var}_B\left[\frac{1}{H}\sum_{h=1}^{H}\frac{1}{m_{Ah}}\sum_{r=1}^{m_{Ah}}\mathbf{1}'\hat{\mathbf{V}}_B\mathbf{1}\right] \\
&\quad + E_B\left[\frac{1}{H^2}\sum_{h=1}^{H}\frac{1}{m_{Ah}}\{2tr(\mathbf{1}\mathbf{1}'\hat{\mathbf{V}}_B\mathbf{1}\mathbf{1}'\hat{\mathbf{V}}_B)\}\right] \\
&= \mathrm{Var}_B[\mathbf{1}'\hat{\mathbf{V}}_B\mathbf{1}] + \frac{2}{H\bar{m}_A^*}[E_B(\mathbf{1}'\hat{\mathbf{V}}_B\mathbf{1})^2],
\end{aligned}
$$

where $\bar{m}_A^* = (H^{-1}\sum_{h=1}^{H} m_{Ah}^{-1})^{-1}$ is the harmonic mean of m_{Ah}.

References

Binder, D.A. (1995). Linearization methods for single phase and two-phase samples: A cookbook approach. *Survey Methodology*, 22, 1, 17-22.

Bray, R., Hourani, L., Rae, K., Dever, J., Brown, J., Vincus, A., Pemberton, M., Marsden, M., Faulkner, D. and Vandermaas-Peeler, R. (2003). 2002 Department of Defense Survey of Health Related Behaviors Among Military Personnel. Tech. Rep. RTI/7841/006-FR, U.S. Department of Defense prepared by RTI International. URL http://dodwws.rti.org/2002WWFinalReportComplete05-04.pdf.

Canty, A.J., and Davison, A.C. (1999). Resampling-based variance estimation for Labour Force Surveys. *The Statistician*, 48, 379-391.

Centers for Disease Control and Prevention (2006). Technical Information and Data for the Behavioral Risk Factor Surveillance System (BRFSS) – BRFSS Weighting Formula. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, September 11, 2006.

Demnati, A., and Rao, J.N.K. (2004). Linearization variance estimators for survey data. *Survey Methodology*, 30, 1, 17-26.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418), 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(423), 1013-1020.

Estevao, V.M., and Särndal, C.-E. (2000). A Functional form approach to calibration. *Journal of Official Statistics*, 16(4), 379-399.

Fuller, W.A. (1998). Replication variance estimation for the two-phase samples. *Statistica Sinica*, 8, 1153-1164.

Hidiroglou, M.A., and Patak, Z. (2006). Raking ratio estimation: An application to the Canadian Retail Trade Survey. *Journal of Official Statistics*, 22(1), 71-80.

Isaki, C.T., Tsay, J.H. and Fuller, W.A. (2004). Weighting sample data subject to independent controls. *Survey Methodology*, 30, 1, 35-44.

Jayasuriya, B.R., and Valliant, R. (1996). An application of regression and calibration estimation to post-stratification in a Household Survey. *Survey Methodology*, 22, 2, 127-137.

Keeter, S., Dimock, M. and Christian, L. (2008). Calling Cell Phones in '08 Pre-Election Polls. NEWS Release (December 18, 2008): Pew Research Center for the People & the Press. URL http://people-press.org/reports/pdf/cell-phone-commentary.pdf.

Killion, R.A. (2006). Weighting Specifications for The American Time Use Survey (ATUS) for 2006. U.S. Bureau of the Census, Internal Memo (Doc.#ATUS-16).

Kim, J.J., Li, J. and Valliant R. (2007). Cell collapsing in poststratification. *Survey Methodology*, 33, 2, 139-150.

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

Krewski, D., and Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, 9(5), 1010-1019.

Krotki, K. (2007). Combining RDD and Web Panel Surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association (in print).

Lessler, J.T., and Kalsbeek, W.D. (1992). *Nonsampling Error in Surveys*. New York: John Wiley & Sons, Inc.

Lumley, T. (2009). Survey: Analysis of complex survey samples. R package version 3.19. University of Washington: Seattle.

Mirza, H., and Hörngren, J. (2002). The Sampling and the Estimation Procedure in the Swedish Labour Force Survey. Technical report, Statistics Sweden, Stockholm: Sweden.

Nadimpalli, V., Judkins, D. and Chu, A. (2004). Survey Calibration to CPS Household Statistics. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 4090-4094.

R Development Core Team (2009). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available: http://www.R-project.org.

Rao, J.N.K., and Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86(2), 403-415.

Rao, J.N.K., and Wu, C.F.J. (1985). Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80(391), 620-630.

Research Triangle Institute (2008). *SUDAAN Language Manual*. Release 10.0, Research Triangle Park, NC: Research Triangle Institute.

Rust, K.F., and Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.

Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. England: John Wiley & Sons, Inc.

Särndal, C.-E., Swensson, B. and Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76(3), 527-537.

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

SAS Institute Inc. (2009). *SAS/STAT® 9.2 User's Guide*. Cary, NC: SAS Institute Inc.

Searle, S.R. (1982). *Matrix Algebra Useful for Statistics*. New York: John Wiley & Sons, Inc.

StataCorp (2010). *Stata Statistical Software: Release 11*. Survey Data, College Station, TX: StataCorp LP.

Stukel, D.M., Hidiroglou, M.A. and Särndal, C.-E. (1996). Variance estimation for calibration estimators: A comparison of jackknifing versus Taylor linearization. *Survey Methodology*, 22, 2, 117-125.

Taylor, M.F., Brice, J., Buck, N. and Prentice-Lane, E. (2007). *British Household Panel Survey User Manual Volume A*: *Introduction, Technical Report and Appendices*. University of Essex, Colchester.

Terhanian., G., Bremer, J., Smith, R. and Thomas, R. (2000). *Correcting Data from Online Survey for the Effects of Nonrandom Selection and Nonrandom Assignment*. Research Paper: Harris Interactive.

Thĕberge, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94(446), 635-644.

Valliant, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, 88, 89-96.

Wolter, K.M. (2007). *Introduction to Variance Estimation*. New York: Springer Science+Business Media, LLC.

Wu, C.F.J. (1985). Variance estimation for the combined ratio and combined regression estimators. *Journal of the Royal Statistical Society*, Series B, 47(1), 147-154.

Yung, W., and Rao, J.N.K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology*, 22, 23-31.

Yung, W., and Rao, J.N.K. (2000). Jackknife variance estimation under imputation for estimators using poststratification information. *Journal of the American Statistical Association*, 95(451), 903-915.

Survey Methodology, June 2010
Vol. 36, No. 1, pp. 57-68
Statistics Canada, Catalogue No. 12-001-X

Some contributions to jackknifing two-phase sampling estimators

Patrick J. Farrell and Sarjinder Singh [1]

Abstract

In this paper, the problem of estimating the variance of various estimators of the population mean in two-phase sampling has been considered by jackknifing the two-phase calibrated weights of Hidiroglou and Särndal (1995, 1998). Several estimators of population mean available in the literature are shown to be the special cases of the technique developed here, including those suggested by Rao and Sitter (1995) and Sitter (1997). By following Raj (1965) and Srivenkataramana and Tracy (1989), some new estimators of the population mean are introduced and their variances are estimated through the proposed jackknife procedure. The variance of the chain ratio and regression type estimators due to Chand (1975) are also estimated using the jackknife. A simulation study is conducted to assess the efficiency of the proposed jackknife estimators relative to the usual estimators of variance.

Key Words: Auxiliary information; Calibration; Estimation of mean and variance; Jackknife; Two-phase sampling.

1. Introduction

Hidiroglou and Särndal (1995, 1998) have pointed out that two-phase sampling for the estimation of finite population attributes is a powerful and cost-effective technique, and hence plays an eminent role in survey sampling. Two-phase sampling can be described as follows. Consider a finite population that we shall denote by $\Omega = \{1, 2, \ldots, i, \ldots, N\}$. Suppose that information is available on a variable Z across the entire population; that is, the values Z_i for all $i = 1, \ldots, N$, are known, implying that the population mean, $\bar{Z}$, is also known. A first-phase probability sample s_1, $s_1 \subset \Omega$, of size m is drawn from the population with selection probabilities π_{1i}. Thus, the first-phase sampling weights can be defined as $d_{1i} = 1/\pi_{1i}$. Assume that for this sample, information is collected on a variable X, which is then paired with the information on Z for each of the m units, giving rise to the data $\{(x_i, z_i) \mid i \in s_1\}$ for $i = 1, \ldots, m$. Once the first-phase sample s_1 has been drawn, a second-phase sample s_2, $s_2 \subset s_1 \subset \Omega$, of size n is selected from s_1 with selection probabilities $\pi_{2i} = \pi_{i|s_1}$, allowing for the second-phase sampling weights to be defined as $d_{2i} = 1/\pi_{2i}$. In the second-phase sample, information is now collected on a variable Y for each selected unit. This information is linked to that previously available on Z and X for these units, giving rise to the data $\{(x_i, y_i, z_i) \mid i \in s_2\}$ for $i = 1, \ldots, n$. Suppose that interest lies in estimating the population mean $\bar{Y}$, and on the variance of the estimator employed.

Let $w_{1i}^o = d_{1i} / \sum_{i \in s_1} d_{1i}$ denote the first-phase normalized original design weights. The usual estimator of the population mean $\bar{X}$ is given by

$$\hat{\bar{X}}_1^o = \sum_{i \in s_1} w_{1i}^o x_i,$$

while a calibrated first-phase estimator of $\bar{X}$ is

$$\hat{\bar{X}}_1^c = \sum_{i \in s_1} w_{1i}^c x_i,$$

where the w_{1i}^c are calibrated weights such that the chi-square distance function

$$D_1 = \sum_{i \in s_1} \{(w_{1i}^c - w_{1i}^o)^2 / (w_{1i}^o q_{1i})\}, \qquad (1.1)$$

is minimized subject to

$$\sum_{i \in s_1} w_{1i}^c z_i = \bar{Z}. \qquad (1.2)$$

In (1.1), the q_{1i} are a set of suitably chosen weights. Minimization of (1.1) subject to (1.2) leads to the first-phase calibrated weights

$$w_{1i}^c = w_{1i}^o + \left\{(q_{1i} w_{1i}^o z_i) \Big/ \left(\sum_{i \in s_1} q_{1i} w_{1i}^o z_i^2\right)\right\}\left(\bar{Z} - \sum_{i \in s_1} w_{1i}^o z_i\right).$$

Thus, a first-phase calibrated estimator of $\bar{X}$ is given by

$$\hat{\bar{X}}_1^c = \sum_{i \in s_1} w_{1i}^o x_i + \hat{\beta}_1 \left(\bar{Z} - \sum_{i \in s_1} w_{1i}^o z_i\right),$$

where

$$\hat{\beta}_1 = \left(\sum_{i \in s_1} q_{1i} w_{1i}^o x_i z_i\right) \Big/ \left(\sum_{i \in s_1} q_{1i} w_{1i}^o z_i^2\right).$$

Now, let $w_{2i}^o = d_{1i} d_{2i} / \sum_{i \in s_2} d_{1i} d_{2i}$ denote the second-phase normalized design weights. The usual estimator of $\bar{Y}$ is given by

1. Patrick J. Farrell, School of Mathematics and Statistics, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario, Canada, K1S 5B6. E-mail: pfarrell@math.carleton.ca; Sarjinder Singh, Department of Mathematics, Texas A&M University - Kingsville, Kingsville, Texas, U.S.A., 78363. E-mail: sarjinder@yahoo.com.

$$\hat{\bar{Y}}_2^o = \sum_{i \in s_2} w_{2i}^o y_i.$$

Let us consider the second-phase calibrated estimator of $\bar{Y}$ as

$$\hat{\bar{Y}}^c = \sum_{i \in s_2} w_{2i}^c y_i, \quad (1.3)$$

where the w_{2i}^c are the second-phase calibrated weights such that the chi-square distance function

$$D_2 = \sum_{i \in s_2} \{(w_{2i}^c - w_{2i}^o)^2 / (w_{2i}^o q_{2i})\}, \quad (1.4)$$

is minimized subject to the calibration constraint

$$\sum_{i \in s_2} w_{2i}^c x_i = \hat{\bar{X}}_1^c. \quad (1.5)$$

Minimization of (1.4) subject to (1.5) leads to the second-phase calibrated weights

$$w_{2i}^c = w_{2i}^o + \left\{(q_{2i} w_{2i}^o x_i) \Big/ \left(\sum_{i \in s_2} q_{2i} w_{2i}^o x_i^2\right)\right\} \left(\hat{\bar{X}}_1^c - \sum_{i \in s_2} w_{2i}^o x_i\right).$$

Thus, the second-phase calibrated estimator of $\bar{Y}$ specified in (1.3) can be written as

$$\hat{\bar{Y}}^c = \hat{\bar{Y}}_2^o + \hat{\beta}_2(\hat{\bar{X}}_1^o - \hat{\bar{X}}_2^o) + \hat{\beta}_1 \hat{\beta}_2(\bar{Z} - \hat{\bar{Z}}_1^o), \quad (1.6)$$

where $\hat{\bar{Z}}_1^o = \sum_{i \in s_1} w_{1i}^o z_i$, $\hat{\bar{X}}_1^o = \sum_{i \in s_1} w_{1i}^o x_i$, $\hat{\bar{X}}_2^o = \sum_{i \in s_2} w_{2i}^o x_i$, $\hat{\bar{Y}}_2^o = \sum_{i \in s_2} w_{2i}^o y_i$, and

$$\hat{\beta}_2 = \left(\sum_{i \in s_2} q_{2i} w_{2i}^o x_i y_i\right) \Big/ \left(\sum_{i \in s_2} q_{2i} w_{2i}^o x_i^2\right).$$

Hidiroglou and Särndal (1995, 1998) and Singh (2000) have considered the problem of estimating the variance of the calibrated estimator $\hat{\bar{Y}}^c$ in (1.6) by using a design-based approach. In a more general context, Rao and Sitter (1995) and Sitter (1997) have pointed out that under simple random sampling without replacement (SRSWOR), a jackknife technique can be used to estimate the variances of the ratio and regression estimators for a population mean. These authors have also reported that the use of the jackknife for estimating variance is more convenient and efficient than the traditional techniques based on estimates of moments.

Of late, a number of authors have investigated the use of jackknife procedures for estimating variances (See Arnab and Singh 2006, Berger 2007, Berger and Skinner 2005, Chen and Shao 2001, and Kovar and Chen 1994). Fuller (1998), Kim, Navarro and Fuller (2000, 2006), Kim and Sitter (2003), and Kott and Stukel (1997) have suggested an approach for estimating the variance in two-stage sampling. Fuller (1998) and Kim and Sitter (2003) address the regression estimator. In particular, consider the generalized regression estimator of population total

$$\hat{Y}_{\mathrm{DS}} = \sum_{i \in s_2} \alpha_i y_i,$$

due to Deville and Särndal (1992). Following Kim *et al.* (2000, 2006), for each $k \in s_2$, specify the jackknife estimator of population total as

$$\hat{Y}_{\mathrm{Kim}} = \sum_{i \in s_2 \setminus k} \alpha_i^{(k)} y_i, \quad (1.7)$$

and the chi-square distance between the design and calibration weights as

$$D_{(k)} = (1/2) \sum_{i \in s_2 \setminus k} \{(\alpha_i^{(k)} - w_i^{(k)} w_i^{*(k)})^2 / (w_i^{(k)} q_i^{(k)})\}. \quad (1.8)$$

Minimizing (1.8) subject to the condition

$$\sum_{i \in s_2 \setminus k} \alpha_i^{(k)} x_i = \sum_{i \in s_1 \setminus k} w_i^{(k)} x_i,$$

leads to jackknifed calibrated weights given by

$$\alpha_i^{(k)} = w_i^{(k)} w_i^{*(k)} + \left\{(w_i^{(k)} q_i^{(k)} x_i) \Big/ \left(\sum_{i \in s_2 \setminus k} w_i^{(k)} q_i^{(k)}\right)\right\} \left\{\sum_{i \in s_2 \setminus k} w_i^{(k)} x_i - \sum_{i \in s_2 \setminus k} w_i^{(k)} w_i^{*(k)} x_i\right\}.$$

It would appear that Kim *et al.* (2006) readjusted these weights as

$$\alpha_i^{(k)} = \begin{cases} \alpha_i^{(k)} & \text{if } k \in s_2 \\ w_i^{(k)} & \text{if } j \in (s_1 - s_2). \end{cases}$$

For such a readjustment, the estimator in (1.7) is equivalent to that of Rao and Sitter (1995).

In the present paper, we consider a new jackknife technique to estimate the variance of the estimator $\hat{\bar{Y}}^c$ under the two-phase setup by following Hidiroglou and Särndal (1995, 1998). Similar to Kim *et al.* (2006), the estimator proposed by Rao and Sitter (1995) is shown to be a special case of the proposed method. However, our approach differs from that of Fuller (1998) Kim and Sitter (2003), Kim *et al.* (2000, 2006) in that we consider calibration at both the first and second phases, thus allowing for the development of the technique for chain ratio and chain regression type estimators. We also investigate, via a simulation study, the efficiency of the jackknife estimators of variance relative to the usual estimators.

2. Estimation of variance using jackknifing

In what follows, we assume that a single stage design is employed at both of the two phases in the sampling process.

Let $\hat{\bar{Y}}^c(j)$ be a calibrated estimator of the population mean, $\bar{Y}$, obtained by dropping the j^{th} unit from the sample s_1 of m units. We prove in the Appendix that the jackknife estimator of the population mean in two phase-sampling can be written as

$$\hat{\bar{Y}}^c(j) = \begin{cases} \hat{\bar{Y}}_2^o(j) + \hat{\beta}_2(j)\,\{\hat{\bar{X}}_1^o(j) - \hat{\bar{X}}_2^o(j)\} \\ \quad + \hat{\beta}_1(j)\,\hat{\beta}_2(j)\{\bar{Z} - \hat{\bar{Z}}_1^o(j)\} & \text{if } j \in s_2 \\ \\ \hat{\bar{Y}}_2^o + \hat{\beta}_2\,\{\hat{\bar{X}}_1^o(j) - \hat{\bar{X}}_1^o\} \\ \quad + \hat{\beta}_1(j)\,\hat{\beta}_2\{\bar{Z} - \hat{\bar{Z}}_1^o(j)\} & \text{if } j \in (s_1 - s_2) \end{cases} \tag{2.1}$$

where the quantity $\hat{\bar{Z}}_1^o(j) = \hat{\bar{Z}}_1^o + \{w_{1j}^o/(1 - w_{1j}^o)\}\{\hat{\bar{Z}}_1^o - z_j\}$, the terms $\hat{\bar{X}}_1^o(j)$, $\hat{\bar{X}}_2^o(j)$, and $\hat{\bar{Y}}_2^o(j)$ are defined in an analogous manner, $\hat{\beta}_1(j) = \hat{\beta}_1 + \{q_{1j} w_{1j}^o z_j (x_j - \hat{\beta}_1 z_j)\}/\{q_{1j} w_{1j}^o z_j^2 - \sum_{i \in s_1} q_{1i} w_{1i}^o z_i^2\}$, and $\hat{\beta}_2(j) = \hat{\beta}_2 + \{q_{2j} w_{2j}^o x_j (y_j - \hat{\beta}_2 x_j)\}/\{q_{2j} w_{2j}^o x_j^2 - \sum_{i \in s_1} q_{2i} w_{2i}^o x_i^2\}$. The modified jackknife estimator of variance of $\hat{\bar{Y}}^c$ is then given by

$$\hat{V}_{\text{JACK}}(\hat{\bar{Y}}^c) = \{(m-1)/m\} \sum_{j \in s_1} \{\hat{\bar{Y}}^c(j) - \hat{\bar{Y}}^c\}^2. \tag{2.2}$$

We show in the appendix that this estimator is consistent.

Note that we can write that

$$\hat{\bar{Y}}^c(j) - \hat{\bar{Y}}^c = \begin{cases} \varepsilon_2(j) + \hat{\beta}_2 \varepsilon_1(j) + \hat{\beta}_2(j)\, d_2(j) \\ \quad + \hat{\beta}_2 \delta_2(j) & \text{if } j \in s_2 \\ \\ \hat{\beta}_2 \varepsilon_1(j) & \text{if } j \in (s_1 - s_2) \end{cases} \tag{2.3}$$

where the terms in (2.3) are given by $\varepsilon_1(j) = \{\hat{\bar{X}}_1^o(j) - \hat{\bar{X}}_1^o\} - \hat{\beta}_1(j)\{\hat{\bar{Z}}_1^o(j) - \bar{Z}\}$, $\varepsilon_2(j) = \{\hat{\bar{Y}}_2^o(j) - \hat{\bar{Y}}_2^o\} - \hat{\beta}_2(j)\{\hat{\bar{X}}_2^o(j) - \hat{\bar{X}}_2^o\} - \hat{\beta}_1(j)\,\hat{\beta}_2(j)\{\hat{\bar{Z}}_1^o(j) - \bar{Z}\}$, $d_2(j) = \{\hat{\bar{X}}_1^o(j) - \hat{\bar{X}}_2^o(j)\}$ and $\delta_2(j) = \{\hat{\bar{X}}_2^o(j) - \hat{\bar{X}}_1^o\} - \hat{\beta}_1(j)\{\bar{Z} - \hat{\bar{Z}}_1^o(j)\} - \hat{\beta}_1\{\bar{Z} - \hat{\bar{Z}}_1^o\}$. The $\varepsilon_1(j)$ term is analogous to the error term associated with the regression of the auxiliary variable x_i on z_i, for $i \in s_1$, while $\varepsilon_2(j)$ is analogous to the error term associated with the regression of the study variable y_i on both x_i and z_i simultaneously, for $i \in s_2$. Provided that $j \in s_2$, the $d_2(j)$ term reflects the difference in the jackknife first and second phase sample means for the variable X, while $\delta_2(j)$ denotes an adjustment to $d_2(j)$ obtained by using information on the auxiliary variable Z.

Using (2.3) in (2.2), the jackknife estimator of variance of the estimator $\hat{\bar{Y}}^c$ is given by

$$\begin{aligned} \hat{V}_{\text{JACK}}(\hat{\bar{Y}}^c) = \{(m-1)/m\} \Bigg[& \sum_{j \in s_2} \varepsilon_2^2(j) + \sum_{j \in s_2} \hat{\beta}_2^2(j) d_2^2(j) \\ & + \hat{\beta}_2^2 \sum_{j \in s_2} \delta_2(j)\{\delta_2(j) + 2\varepsilon_1(j)\} \\ & + 2\hat{\beta}_2 \sum_{j \in s_2} \varepsilon_1(j)\varepsilon_2(j) \\ & + 2\hat{\beta}_2 \sum_{j \in s_2} \hat{\beta}_2(j)\, d_2(j)\{\varepsilon_1(j) + \delta_2(j)\} \\ & + \hat{\beta}_2^2 \sum_{j \in s_1} \varepsilon_1^2(j) \Bigg]. \end{aligned} \tag{2.4}$$

Note that the expression given in (2.4) is exact. It can be used to estimate the variance of several estimators available in the literature.

3. Special cases

In the next section, we demonstrate that the estimator proposed by Rao and Sitter (1995), Sitter (1997), Raj (1965), Srivenkataramana and Tracy (1989), Chand (1975), and Ahmed (1997) can be viewed as special cases of the proposed technique.

Case 3.1: Rao and Sitter (1995)

If $\hat{\bar{X}}_1^c = \hat{\bar{X}}_1^o$ (no first-phase calibration is made) and $q_{2i} = 1/x_i$, then the calibrated estimator of $\bar{Y}$ becomes

$$\hat{\bar{Y}}_r^c = \left(\sum_{i \in s_2} w_{2i}^o y_i\right) \left\{ \left(\sum_{i \in s_2} w_{1i}^o x_i\right) \Big/ \left(\sum_{i \in s_2} w_{2i}^o x_i\right) \right\}.$$

If the first-phase sample s_1 is selected according to SRSWOR such that the first-phase design weights are given by $d_{1i} = N/m$, and the second-phase sample s_2 is selected from s_1 by SRSWOR such that $d_{2i} = m/n$, then the calibrated estimator of the population mean becomes

$$\hat{\bar{Y}}_{\text{RS}}^c = \bar{y}(\bar{x}'/\bar{x}), \tag{3.1}$$

where $\bar{y} = \sum_{i \in s_2} y_i/n$, $\bar{x} = \sum_{i \in s_2} x_i/n$, and $\bar{x}' = \sum_{i \in s_1} x_i/m$. The jackknife mechanism in (2.1) becomes

$$\hat{\bar{Y}}_{\text{RS}}^c(j) = \begin{cases} \dfrac{(n\bar{y} - y_j)(m\bar{x}' - x_j)}{(n\bar{x} - x_j)(m-1)} & \text{if } j \in s_2 \\ \\ (\bar{y}/\bar{x})\dfrac{(m\bar{x}' - x_j)}{(m-1)} & \text{if } j \in (s_1 - s_2). \end{cases} \tag{3.2}$$

Setting $\hat{R} = \bar{y}/\bar{x}$, the difference between (3.2) and (3.1) can be written as

$$\hat{\bar{Y}}^c_{\mathrm{RS}}(j) - \hat{\bar{Y}}^c_{\mathrm{RS}} =$$

$$\begin{cases} -\hat{R}\dfrac{(x_j - \bar{x}')}{(m-1)} - \dfrac{\bar{x}'(j)}{\bar{x}(j)}\dfrac{(y_j - \hat{R}x_j)}{(n-1)} & \text{if } j \in s_2 \\ -\hat{R}\dfrac{(x_j - \bar{x}')}{(m-1)} & \text{if } j \in (s_1 - s_2). \end{cases} \tag{3.3}$$

Expression (3.3) is exactly the same as reported by Rao and Sitter (1995). Assuming that $\bar{x}'(j)/\bar{x}(j) \approx \bar{x}'/\bar{x}$, then the approximate jackknife estimator of variance is given by

$$\hat{V}_{\mathrm{JACK}}(\hat{\bar{Y}}^c_{\mathrm{RS}}) \approx \left(\frac{\bar{x}'}{\bar{x}}\right)^2 \sum_{i\in s_2} \frac{(y_i - \hat{R}x_i)^2}{n(n-1)} + 2\left(\frac{\bar{x}'}{\bar{x}}\right)\hat{R}\sum_{j\in s_2}\frac{(x_j - \bar{x}')(y_j - \hat{R}x_j)}{n-1} + \hat{R}^2\sum_{j\in s_1}\frac{(x_j - \bar{x}')^2}{m(m-1)}.$$

Thus, the Rao and Sitter (1995) estimator is a special case of the proposed jackknife technique.

Case 3.2: Sitter (1997)

In Case 3.1, if we consider $q_{2i} = 1$, then the calibrated estimator under SRSWOR becomes

$$\hat{\bar{Y}}^c_{lr} = \bar{y} + b^*(\bar{x}' - \bar{x}), \tag{3.4}$$

where $b^* = \sum_{i\in s_2} x_i y_i / \sum_{i\in s_2} x_i^2$ denotes an estimator of the regression coefficient β that is slightly different from the one considered by Sitter (1997). The jackknife mechanism takes the form

$$\hat{\bar{Y}}^c_{lr}(j) = \begin{cases} \dfrac{n\bar{y} - y_j}{n-1} + \left\{ b^* + \dfrac{x_j(y_j - b^* x_j)}{\sum_{i\in s_2} x_i^2 - x_j^2} \right\} \left\{ \dfrac{m\bar{x}' - x_j}{m-1} - \dfrac{n\bar{x} - x_j}{n-1} \right\} & \text{if } j \in s_2 \\ \bar{y} + b^*\left\{ \dfrac{m\bar{x}' - x_j}{m-1} - \bar{x} \right\} & \text{if } j \in (s_1 - s_2). \end{cases} \tag{3.5}$$

If we set $d_j^* = (y_j - \bar{y}) - b^*(x_j - \bar{x})$, $a_j^* = x_j\{\bar{x}(j) - \bar{x}'(j)\}/K$, and $k_j^* = x_j^2/K$, where $K = (n-1)s_x^2 + n\bar{x}^2$, then the difference between (3.5) and (3.4) can be written as

$$\hat{\bar{Y}}^c_{lr}(j) - \hat{\bar{Y}}^c_{lr} =$$

$$\begin{cases} -b^*\dfrac{(x_j - \bar{x}')}{(m-1)} - \dfrac{\dfrac{d_j^*}{(n-1)}}{\left[1 + \dfrac{a_j^*}{(1-k_j^*)}\right]} & \text{if } j \in s_2 \\ -b^*\dfrac{(x_j - \bar{x}')}{(m-1)} & \text{if } j \in (s_1 - s_2) \end{cases}$$

which is similar to the expression reported by Sitter (1997).

Case 3.3: Raj (1965)

In order to consider this case, we assume that the initial sample s_1 of size m is selected with replacement according to probabilities p_i proportional to z_i, $i = 1, 2, ..., N$. Information on the auxiliary variable X is collected on this first-phase sample, s_1. The second-phase sample, specified to be of size n, is a subsample of s_1 selected without replacement using equal probabilities. It is for s_2 that information on Y is collected. Under this sampling scheme, $d_{1i} = 1/\pi_{1i} = 1/(mp_i)$ and $d_{2i} = m/n$. Thus, $w_{1i}^o = (1/p_i)/\sum_{i\in s_1}(1/p_i)$ and $w_{2i}^o = (1/p_i)/\sum_{i\in s_2}(1/p_i)$. Note also that for this scheme, $\hat{\bar{X}}_1^c = \hat{\bar{X}}_1^o$; thus no first-phase calibration is made. If $q_{2i} = 1/x_i$, then the calibrated estimator $\hat{\bar{Y}}^c$ becomes

$$\hat{\bar{Y}}^c_{\mathrm{Raj}} = \hat{\bar{Y}}_2^o(\hat{\bar{X}}_1^o / \hat{\bar{X}}_2^o), \tag{3.6}$$

where $\hat{\bar{Y}}_2^o = \sum_{i\in s_2}(y_i/p_i)/\sum_{i\in s_2}(1/p_i)$, $\hat{\bar{X}}_2^o = \sum_{i\in s_2}(x_i/p_i)/\sum_{i\in s_2}(1/p_i)$, and $\hat{\bar{X}}_1^o = \sum_{i\in s_1}(x_i/p_i)/\sum_{i\in s_1}(1/p_i)$. Thus, alternatively $\hat{\bar{Y}}^c_{\mathrm{Raj}} = \{\sum_{i\in s_2}(y_i/p_i)\sum_{i\in s_1}(x_i/p_i)\}/\{\sum_{i\in s_2}(x_i/p_i)\sum_{i\in s_1}(1/p_i)\}$.

Under the sampling scheme described above, the jackknife estimator of population mean is

$$\hat{\bar{Y}}^c_{\mathrm{Raj}}(j) = \begin{cases} \hat{\bar{Y}}_2^o(j)\dfrac{\hat{\bar{X}}_1^o(j)}{\hat{\bar{X}}_2^o(j)} & \text{if } j \in s_2 \\ \hat{\bar{Y}}_2^o\dfrac{\hat{\bar{X}}_1^o(j)}{\hat{\bar{X}}_2^o} & \text{if } j \in (s_1 - s_2) \end{cases} \tag{3.7}$$

where

$$\hat{\bar{Y}}_2^o(j) = \frac{\sum_{i\in s_2}(y_i/p_i)}{\sum_{i\in s_2}(1/p_i)} + \frac{(1/p_j)\Big/\sum_{i\in s_2}(1/p_i)}{1 - \dfrac{(1/p_j)}{\sum_{i\in s_2}(1/p_i)}}\left\{\frac{\sum_{i\in s_2}(y_i/p_i)}{\sum_{i\in s_2}(1/p_i)} - y_j\right\},$$

and $\hat{\bar{X}}_2^o(j)$ and $\hat{\bar{X}}_1^o(j)$ are defined analogously. If $\hat{R} = \hat{\bar{Y}}_2^o / \hat{\bar{X}}_2^o$ and $w_{2j}^o = (1/p_j)/\sum_{i\in s_2}(1/p_i)$, the difference between (3.7) and (3.6) can easily be written as

$$\hat{\bar{Y}}_{\text{Raj}}^c(j) - \hat{\bar{Y}}_{\text{Raj}}^c = \begin{cases} -w_{2j}^o \dfrac{\hat{\bar{X}}_1^o(j)}{\hat{\bar{X}}_2^o(j)}(y_j - \hat{R}x_j) \\ \qquad + \hat{R}\{\hat{\bar{X}}_1^o(j) - \hat{\bar{X}}_1^o\} & \text{if } j \in s_2 \\ \hat{R}\{\hat{\bar{X}}_1^o(j) - \hat{\bar{X}}_1^o\} & \text{if } j \in (s_1 - s_2). \end{cases}$$

Thus, the jackknife estimator of variance of the estimator $\hat{\bar{Y}}_{\text{Raj}}^c$ is given by

$$\hat{V}_{\text{JACK}}(\hat{\bar{Y}}_{\text{Raj}}^c) = \frac{m-1}{m}\Bigg[\sum_{j\in s_2}(w_{2j}^o)^2 \frac{\hat{\bar{X}}_1^o(j)^2}{\hat{\bar{X}}_2^o(j)}(y_j - \hat{R}x_j)^2 + \hat{R}^2\sum_{j\in s_1}\{\hat{\bar{X}}_1^o(j) - \hat{\bar{X}}_1^o\}^2 - 2\hat{R}\sum_{j\in s_2} w_{2j}^o \frac{\hat{\bar{X}}_1^o(j)}{\hat{\bar{X}}_2^o(j)}(y_j - \hat{R}x_j)\{\hat{\bar{X}}_1^o(j) - \hat{\bar{X}}_1^o\}\Bigg].$$

Following Rao and Sitter (1995), if we assume $\hat{\bar{X}}_1^o(j)/\hat{\bar{X}}_2^o(j) \approx \hat{\bar{X}}_1^o/\hat{\bar{X}}_2^o$, then the jackknife estimator of variance of $\hat{\bar{Y}}_{\text{Raj}}^c$ takes the form

$$\hat{V}_{\text{JACK}}(\hat{\bar{Y}}_{\text{Raj}}^c) \approx \frac{m-1}{m}\Bigg[\{\hat{\bar{X}}_1^o/\hat{\bar{X}}_2^o\}^2\sum_{j\in s_2}(w_{2j}^o)^2(y_j - \hat{R}x_j)^2 + \hat{R}^2\sum_{j\in s_1}\{\hat{\bar{X}}_1^o(j) - \hat{\bar{X}}_1^o\}^2 - 2\hat{R}\{\hat{\bar{X}}_1^o/\hat{\bar{X}}_2^o\}\sum_{j\in s_2} w_{2j}^o(y_j - \hat{R}x_j)\{\hat{\bar{X}}_1^o(j) - \hat{\bar{X}}_1^o\}\Bigg].$$

Case 3.4: Srivenkataramana and Tracy (1989)

In order to consider this case, as in Raj (1965), we assume that the initial sample s_1 of size m is selected with replacement according to probabilities proportional to z_i. However, the subsample, s_2, of n units is now selected with replacement using probabilities proportional to x_i/z_i. As a result, $w_{1i}^o = (1/z_i)/\sum_{i\in s_1}(1/z_i)$ and $w_{2i}^o = (1/x_i)/\sum_{i\in s_2}(1/x_i)$. Similar to Raj (1965), no first-phase calibration is made; thus $\hat{\bar{X}}_1^c = \hat{\bar{X}}_1^o$. Hence, if $q_{2i} = 1/x_i$, then the calibrated estimator $\hat{\bar{Y}}^c$ is

$$\hat{\bar{Y}}_{\text{ST}}^c = \hat{\bar{Y}}_2^o(\hat{\bar{X}}_1^o/\hat{\bar{X}}_2^o), \tag{3.8}$$

where $\hat{\bar{Y}}_2^o = \sum_{i\in s_2}(y_i/x_i)/\sum_{i\in s_2}(1/x_i)$, $\hat{\bar{X}}_2^o = n/\sum_{i\in s_2}(1/x_i)$, and $\hat{\bar{X}}_1^o = \sum_{i\in s_1}(x_i/z_i)/\sum_{i\in s_1}(1/z_i)$. Thus, alternatively $\hat{\bar{Y}}_{\text{ST}}^c = \{\sum_{i\in s_2}(y_i/x_i)\sum_{i\in s_1}(x_i/z_i)\}/\{n\sum_{i\in s_1}(1/z_i)\}$.

Under the sampling scheme described above, the jackknife estimator of population mean is

$$\hat{\bar{Y}}_{\text{ST}}^c(j) = \begin{cases} \hat{\bar{Y}}_2^o(j)\{\hat{\bar{X}}_1^o(j)/\hat{\bar{X}}_2^o(j)\} & \text{if } j \in s_2 \\ \hat{\bar{Y}}_2^o\{\hat{\bar{X}}_1^o(j)/\hat{\bar{X}}_2^o\} & \text{if } j \in (s_1 - s_2) \end{cases} \tag{3.9}$$

where

$$\hat{\bar{Y}}_2^o(j) = \frac{\sum_{i\in s_2}(y_i/x_i)}{\sum_{i\in s_2}(1/x_i)} + \frac{1}{x_j\sum_{i\in s_2}(1/x_i) - 1}\left\{\frac{\sum_{i\in s_2}(y_i/x_i)}{\sum_{i\in s_2}(1/x_i)} - y_j\right\}.$$

The terms $\hat{\bar{X}}_2^o(j)$ and $\hat{\bar{X}}_1^o(j)$ are defined similarly; that is

$$\hat{\bar{X}}_2^o(j) = \frac{n}{\sum_{i\in s_2}(1/x_i)} + \frac{1}{x_j\sum_{i\in s_2}(1/x_i) - 1}\left\{\frac{n}{\sum_{i\in s_2}(1/x_i)} - x_j\right\},$$

while $\hat{\bar{X}}_1^o(j)$ can be written as

$$\hat{\bar{X}}_1^o(j) = \frac{\sum_{i\in s_1}(x_i/z_i)}{\sum_{i\in s_1}(1/z_i)} + \frac{1}{x_j\sum_{i\in s_1}(1/x_i) - 1}\left\{\frac{\sum_{i\in s_1}(y_i/x_i)}{\sum_{i\in s_1}(1/x_i)} - x_j\right\}.$$

If $\hat{R} = \sum_{i\in s_2}(y_i/x_i)/n$ and $w_{2j}^o = (1/x_j)/\sum_{i\in s_2}(1/x_i)$, the difference between (3.9) and (3.8) is given by

$$\hat{\bar{Y}}_{\text{ST}}^c(j) - \hat{\bar{Y}}_{\text{ST}}^c = \begin{cases} -w_{2j}^o \dfrac{\hat{\bar{X}}_1^o(j)}{\hat{\bar{X}}_2^o(j)}(y_j - \hat{R}x_j) + \hat{R}\{\hat{\bar{X}}_1^o(j) - \hat{\bar{X}}_1^o\} & \text{if } j \in s_2 \\ \hat{R}\{\hat{\bar{X}}_1^o(j) - \hat{\bar{X}}_1^o\} & \text{if } j \in (s_1 - s_2). \end{cases}$$

Following Rao and Sitter (1995), if we assume $\hat{\bar{X}}_1^o(j)/\hat{\bar{X}}_2^o(j) \approx \hat{\bar{X}}_1^o/\hat{\bar{X}}_2^o$, then the jackknife estimator of variance of $\hat{\bar{Y}}_{\text{ST}}^c$ takes the form

$$\hat{V}_{\text{JACK}}(\hat{\bar{Y}}_{\text{ST}}^c) \approx \frac{m-1}{m}\Bigg[\{\hat{\bar{X}}_1^o/\hat{\bar{X}}_2^o\}^2\sum_{j\in s_2}(w_{2j}^o)^2(y_j - \hat{R}x_j)^2 + \hat{R}^2\sum_{j\in s_1}\{\hat{\bar{X}}_1^o(j) - \hat{\bar{X}}_1^o\}^2 - 2\hat{R}\{\hat{\bar{X}}_1^o/\hat{\bar{X}}_2^o\}\sum_{j\in s_2} w_{2j}^o(y_j - \hat{R}x_j)\{\hat{\bar{X}}_1^o(j) - \hat{\bar{X}}_1^o\}\Bigg].$$

Case 3.5: Chand (1975)

In order to consider this case, the first-phase sample s_1 of size m is selected using SRSWOR, and both auxiliary

variables Z and X are observed on the chosen units. The subsample, s_2, of n units is also selected using SRSWOR. Obviously, $d_{1i} = N/m$ and $d_{2i} = m/n$, so that $w_{1i}^o = 1/m$ and $w_{2i}^o = 1/n$. If $q_{1i} = 1/z_i$ and $q_{2i} = 1/x_i$, then the calibrated estimator $\hat{\bar{Y}}^c$ becomes

$$\hat{\bar{Y}}_{\text{Ch}}^c = \bar{y}(\bar{x}'/\bar{x})(\bar{Z}/\bar{z}'), \tag{3.10}$$

where

$$\bar{y} = \sum_{i\in s_2} y_i/n,\ \bar{x} = \sum_{i\in s_2} x_i/n,\ \bar{x}' = \sum_{i\in s_1} x_i/m,$$

and $\bar{z}' = \sum_{i\in s_1} z_i/m$. The jackknife estimator of $\bar{Y}$ is

$$\hat{\bar{Y}}_{\text{Ch}}^c(j) = \begin{cases} \bar{y}(j)\dfrac{\bar{x}'(j)}{\bar{x}(j)}\dfrac{\bar{Z}}{\bar{z}'(j)} & \text{if } j\in s_2 \\ \bar{y}(j)\dfrac{\bar{x}'(j)}{\bar{x}}\dfrac{\bar{Z}}{\bar{z}'(j)} & \text{if } j\in(s_1-s_2) \end{cases} \tag{3.11}$$

where $\bar{y}(j) = (n\bar{y}-y_j)/(n-1)$, $\bar{x}(j) = (n\bar{x}-x_j)/(n-1)$, $\bar{x}'(j) = (m\bar{x}'-x_j)/(m-1)$, and finally $\bar{z}'(j) = (m\bar{z}'-z_j)/(m-1)$. If we let $\hat{R}_1 = \bar{x}'/\bar{z}'$ (an estimator of $R_1 = \bar{X}/\bar{Z}$) and $\hat{R}_2 = \bar{y}/\bar{x}$ (an estimator of $R_2 = \bar{Y}/\bar{X}$), and similarly, let $\hat{R}_1(j) = \bar{x}'(j)/\bar{z}'(j)$ and $\hat{R}_2(j) = \bar{y}(j)/\bar{x}(j)$, the difference between (3.11) and (3.10) can be written as

$$\hat{\bar{Y}}_{\text{Ch}}^c(j) - \hat{\bar{Y}}_{\text{Ch}}^c = \begin{cases} \varepsilon_2(j)+\hat{R}_2\varepsilon_1(j)+\hat{R}_2(j)d_2(j)+\hat{R}_2\delta_2(j) & \text{if } j\in s_2 \\ \hat{R}_2\varepsilon_1(j) & \text{if } j\in(s_1-s_2) \end{cases} \tag{3.12}$$

where we can write in (3.12) that $\varepsilon_2(j) = \{\bar{y}(j)-\bar{y}\} - \hat{R}_2(j)\{\bar{x}(j)-\bar{x}\} - \hat{R}_1(j)\hat{R}_2(j)\{\bar{z}'(j)-\bar{Z}\}$, $d_2(j) = \{\bar{x}'(j)-\bar{x}'\}$, $\delta_2(j) = \{\bar{x}(j)-\bar{x}'(j)\} - \hat{R}_1(j)\{\bar{Z}-\bar{z}'(j)\} - \hat{R}_1\{\bar{Z}-\bar{z}'\}$, and finally that the term $\varepsilon_1(j) = \{\bar{x}'(j)-\bar{x}'\} - \hat{R}_1(j)\{\bar{z}'(j)-\bar{Z}\}$. Thus the jackknife estimator of variance of the estimator $\hat{\bar{Y}}_{\text{Ch}}^c$ is given by

$$\begin{aligned}\hat{V}_{\text{JACK}}(\hat{\bar{Y}}_{\text{Ch}}^c) = {}& \{(m-1)/m\}\Big[\sum_{j\in s_2}\varepsilon_2^2(j) + \sum_{j\in s_2}\hat{R}_2^2(j)d_2^2(j) \\ & + \hat{R}_2^2\sum_{j\in s_2}\delta_2(j)\{\delta_2(j)+2\varepsilon_1(j)\} \\ & + 2\hat{R}_2\sum_{j\in s_2}\varepsilon_1(j)\varepsilon_2(j) \\ & + 2\hat{R}_2\sum_{j\in s_2}\hat{R}_2(j)d_2(j)\{\varepsilon_1(j)+\delta_2(j)\} \\ & + \hat{R}_2^2\sum_{j\in s_1}\varepsilon_1^2(j)\Big].\end{aligned}$$

Case 3.6: Ahmed (1997)

Consider the same sample design as in Case 3.5. Rather than $q_{1i}=1/z_i$ and $q_{2i}=1/x_i$ as in Chand (1975), we set $q_{1i}=q_{2i}=1$, and $q_{2i}=1/x_i$, then the calibrated estimator reduces to

$$\hat{\bar{Y}}_{\text{Chlr}}^c = \bar{y} + b_2^*(\bar{x}'-\bar{x}) + b_1^*b_2^*(\bar{Z}-\bar{z}'), \tag{3.13}$$

where $b_2^* = \sum_{i\in s_2}x_iy_i/\sum_{i\in s_2}x_i^2$ and $b_1^* = \sum_{i\in s_1}x_iz_i/\sum_{i\in s_1}z_i^2$. Note that (3.13) is a chain regression type estimator similar to Ahmed (1997). Letting $b_2^*(j) = b_2^* + \{x_j(y_j - b_2^*x_j)/(x_j^2-\sum_{i\in s_2}x_i^2)\}$ and $b_1^*(j) = b_1^* + \{z_j(x_j-b_1^*z_j)/(z_j^2-\sum_{i\in s_1}z_i^2)\}$, after jackknifing the estimator $\hat{\bar{Y}}_{\text{Chlr}}^c$ becomes

$$\hat{\bar{Y}}_{\text{Chlr}}^c(j) = \begin{cases} \bar{y}(j) + b_2^*(j)\{\bar{x}'(j)-\bar{x}(j)\} \\ \quad + b_1^*(j)b_2^*(j)\{\bar{Z}-\bar{z}'(j)\} & \text{if } j\in s_2 \\ \bar{y} + b_2^*\{\bar{x}'(j)-\bar{x}\} \\ \quad + b_1^*(j)b_2^*\{\bar{Z}-\bar{z}'(j)\} & \text{if } j\in(s_1-s_2). \end{cases} \tag{3.14}$$

The difference between (3.14) and (3.13) can be written as

$$\hat{\bar{Y}}_{\text{Chlr}}^c(j) - \hat{\bar{Y}}_{\text{Chlr}}^c = \begin{cases} \varepsilon_2(j)+b_2^*\varepsilon_1(j)+b_2^*(j)d_2(j)+b_2^*\delta_2(j) & \text{if } j\in s_2 \\ b_2^*\varepsilon_1(j) & \text{if } j\in(s_1-s_2) \end{cases} \tag{3.15}$$

where we can write in (3.15) that $\varepsilon_2(j) = \{\bar{y}(j) - \bar{y}\} - b_2^*(j)\{\bar{x}(j)-\bar{x}\} - b_1^*(j)\,b_2^*(j)\{\bar{z}'(j)-\bar{Z}\}$, $d_2(j) = \{\bar{x}'(j)-\bar{x}'\}$, $\delta_2(j) = \{\bar{x}(j)-\bar{x}'(j)\} - b_1^*(j)\{\bar{Z}-\bar{z}'(j)\} - b_1^*\{\bar{Z}-\bar{z}'\}$, and finally that the term $\varepsilon_1(j) = \{\bar{x}'(j) - \bar{x}'\} - b_1^*(j)\{\bar{z}'(j)-\bar{Z}\}$. Thus the jackknife estimator of variance of the estimator $\hat{\bar{Y}}_{\text{Chlr}}^c$ is given by

$$\begin{aligned}\hat{V}_{\text{JACK}}(\hat{\bar{Y}}_{\text{Chlr}}^c) = {}& \{(m-1)/m\}\Big[\sum_{j\in s_2}\varepsilon_2^2(j) + \sum_{j\in s_2}\{b_2^*(j)\}^2d_2^2(j) \\ & + \{b_2^*(j)\}^2\sum_{j\in s_2}\delta_2(j)\{\delta_2(j)+2\varepsilon_1(j)\} \\ & + 2b_2^*\sum_{j\in s_2}\varepsilon_1(j)\varepsilon_2(j) \\ & + 2b_2^*\sum_{j\in s_2}b_2^*(j)d_2(j)\{\varepsilon_1(j)+\delta_2(j)\} \\ & + \{b_2^*\}^2\sum_{j\in s_1}\varepsilon_1^2(j)\Big].\end{aligned}$$

4. Simulation study

In this section, we present the results of simulation studies designed to investigate the performance of the proposed jackknife procedure for estimating the variance of four of the two-phase estimators of population mean

presented in Section 3. Specifically, we consider the Rao and Sitter (1995) ratio-type estimator, the Sitter (1997) regression-type estimator, the Chand (1975) chain ratio-type estimator, and the Ahmed (1997) chain regression-type estimator. Initially, we describe and report the results of simulations that were conducted for the Sitter and Rao (1995) and Sitter (1997) estimators. This is followed by a discussion and summary of similar simulations on the Chand (1975) and Ahmed (1997) estimators. Unlike the case for the ratio and regression estimators, since complete information on a second auxiliary variable Z is required for the entire population in order to apply the two chain estimators, the simulations that were conducted for these two estimators are somewhat more complicated than those performed for the ratio and regression estimators.

4.1 Simulation study: Rao and Sitter (1995) and Sitter (1997)

For purposes of the first set of simulations, we assume that a first-phase sample of m units is selected from a population of N units, and only the auxiliary variable X is measured. From the first-phase sample of m units, we then select a second-phase sample of n units by SRSWOR in which both the study variable, Y, and the auxiliary variable, X, are measured.

We began by creating a population of N units consisting of (X_i, Y_i) pairs using the model

$$Y_i = \beta X_i + \sqrt{X_i^g}\,\varepsilon_i,$$

with $\beta = 10$. Initially, we set $g = 0$ and $N = 500$. For each i, $i = 1, ..., N$, we generated X_i from a gamma distribution with a shape parameter of 3.1 and a scale parameter of one, and ε_i from a standard normal. From the resulting population of (X_i, Y_i) pairs, we selected 1,000 first-phase sample of $m = 100$ units, and from each of these samples, we selected 10,000 second-phase samples of $n = 20$ units.

Under the sampling scheme used here, Rao and Sitter (1995) proposed the ratio estimator

$$\hat{\bar{Y}}_{\mathrm{RS}}^{c} = \bar{y}(\bar{x}'/\bar{x}), \tag{4.1}$$

which has approximate variance

$$V(\hat{\bar{Y}}_{\mathrm{RS}}^{c}) = (n^{-1} - m^{-1})S_d^2 + (m^{-1} - N^{-1})S_y^2,$$

where

$$S_d^2 = (N-1)^{-1}\sum_{i=1}^{N}[(Y_i - \bar{Y}) - R(X_i - \bar{X})]^2$$

and

$$S_y^2 = (N-1)^{-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2,$$

with $\bar{Y} = \sum_{i=1}^{N} Y_i/N$, $\bar{X} = \sum_{i=1}^{N} X_i/N$, and $R = \bar{Y}/\bar{X}$. For the t^{th} second phase sample ($t = 1, \ldots, 10{,}000$) drawn from the k^{th} first phase sample ($k = 1, \ldots, 1{,}000$), we computed the usual estimator of variance

$$\hat{V}[(\hat{\bar{Y}}_{\mathrm{RS}}^{c}(t|k)] = \left(\frac{1}{n} - \frac{1}{m}\right)s_{d(t|k)}^2 + \left(\frac{1}{m} - \frac{1}{N}\right)s_{y(t|k)}^2, \tag{4.2}$$

where the sample variances are

$$s_{d(t|k)}^2 = (n-1)^{-1}\sum_{i=1}^{n}[(y_{i(t|k)} - \bar{y}_{(t|k)}) - r_{(t|k)}(x_{i(t|k)} - \bar{x}_{(t|k)})]^2$$

and

$$s_{y(t|k)}^2 = (n-1)^{-1}\sum_{i=1}^{n}(y_{i(t|k)} - \bar{y}_{(t|k)})^2$$

with $\bar{y}_{(t|k)} = \sum_{i=1}^{n} y_{i(t|k)}/n$ and $\bar{x}_{(t|k)} = \sum_{i=1}^{n} x_{i(t|k)}/n$. In addition, $r_{(t|k)} = \bar{y}_{(t|k)}/\bar{x}_{(t|k)}$. We also computed the jackknife estimator of variance

$$\hat{V}_{\mathrm{JACK}}[(\hat{\bar{Y}}_{\mathrm{RS}}^{c}(t|k)] =$$

$$\frac{m-1}{m}\sum_{j=1}^{m}\left[\bar{y}_{(t|k)}(j)\frac{\bar{x}'_{(t|k)}(j)}{\bar{x}_{(t|k)}(j)} - \bar{y}_{(t|k)}\frac{\bar{x}'_{(t|k)}}{\bar{x}_{(t|k)}}\right]^2, \tag{4.3}$$

and the ratio of estimated variances

$$\mathrm{RV}(t|k) = \hat{V}[(\hat{\bar{Y}}_{\mathrm{RS}}^{c}(t|k)]/\hat{V}_{\mathrm{JACK}}[(\hat{\bar{Y}}_{\mathrm{RS}}^{c}(t|k)].$$

We then computed the average of the $\mathrm{RV}(t|k)$ over all k and t, which is given by

$$\mathrm{RV} = \frac{1}{10{,}000{,}000}\sum_{k=1}^{1{,}000}\sum_{t=1}^{10{,}000}\mathrm{RV}(t|k).$$

We also determined empirical estimates of the biases in (4.2) and (4.3) by computing

$$\mathrm{EBU} = \frac{1}{10{,}000{,}000}\sum_{k=1}^{1{,}000}\sum_{t=1}^{10{,}000}\{\hat{V}[(\hat{\bar{Y}}_{\mathrm{RS}}^{c}(t|k)] - V(\hat{\bar{Y}}_{\mathrm{RS}}^{c})\},$$

and

$$\mathrm{EBJ} = \frac{1}{10{,}000{,}000}\sum_{k=1}^{1{,}000}\sum_{t=1}^{10{,}000}\{\hat{V}_{\mathrm{JACK}}[(\hat{\bar{Y}}_{\mathrm{RS}}^{c}(t|k)] - V(\hat{\bar{Y}}_{\mathrm{RS}}^{c})\}.$$

Note that the estimator given in (4.2) is unbiased. Finally, we calculated the relative efficiency of the usual estimator of variance to the jackknife estimator according to

$$\mathrm{RE} = \left(\frac{\sum_{k=1}^{1{,}000}\sum_{t=1}^{10{,}000}\{\hat{V}[(\hat{\bar{Y}}_{\mathrm{RS}}^{c}(t|k)] - V(\hat{\bar{Y}}_{\mathrm{RS}}^{c})\}^2}{\sum_{k=1}^{1{,}000}\sum_{t=1}^{10{,}000}\{\hat{V}_{\mathrm{JACK}}[(\hat{\bar{Y}}_{\mathrm{RS}}^{c}(t|k)] - V(\hat{\bar{Y}}_{\mathrm{RS}}^{c})\}^2}\right).$$

Using the same generated population of $N = 500$, we repeated the simulation; however we used $m = 400$ and $n = 80$ instead. We then created four additional populations of size $N = 500$ using $g = 0.5$, 1.0, 1.5, and 2.0. For each of these four populations, we repeated the two simulations described above where in the first simulation, $m = 100$ with $n = 20$, and in the second simulation, $m = 400$ and $n = 80$. Finally, to study the effect of population size, we then repeated all the simulations based on the different values of g, m, and n when $N = 500$ for three additional values of N, namely 5,000, 50,000, and 500,000. The results obtained for RV, EBU, EBJ, and RE for each of these simulations are presented in Table 1.

The results for RE in Table 1 suggest that as the population size N tends to infinity (as considered by Rao and Sitter 1995), the jackknife estimator of variance remains more efficient than the usual unbiased estimator of variance. It is also the case for very large N that the values for RV tend to one. However, considering the cases where $N = 500$, if the population size is relatively small, not only are the values for RV noticeably smaller than one, but the jackknife estimator of variance seems to be significantly biased. In addition, the jackknife estimator appears to be much less efficient than the usual unbiased estimator of variance, especially when m and n are large. Of note here is the fact that Rao and Sitter (1995) and Sitter (1997) state that it is not clear how to fix the finite population correction factors in the jackknife estimator of variance in two-phase sampling. This would seem to be an area where further research could be fruitful, since it would appear that when the population size is small, it might be worthwhile to adjust the finite population correction factors instead of directly applying the jackknife technique according to the approach proposed here. Note that Kim *et al.* (2006) have incorporated a finite population correction factor in a special case.

Table 1
Comparison of the jackknife and usual estimators of variance of the ratio estimator of the population mean when β = 10 and the auxiliary variable, *X*, follows a gamma distribution with a shape parameter of 3.1 and a scale parameter of one

N	*m*	*n*	*g*	RV	EBU	EBJ	RE
500	100	20	0.0	0.801	0.006	0.542	1.521
			0.5	0.800	0.010	0.579	1.310
			1.0	0.805	-0.071	0.561	1.267
			1.5	0.816	-0.358	0.575	1.149
			2.0	0.840	-0.720	1.777	0.935
5,000	100	20	0.0	0.979	-0.028	0.042	4.015
			0.5	0.976	0.007	0.096	3.709
			1.0	0.965	0.023	0.172	3.210
			1.5	0.936	-0.073	0.337	1.308
			2.0	0.916	-1.103	0.493	0.967
50,000	100	20	0.0	1.001	-0.002	0.003	6.241
			0.5	0.998	0.107	0.126	4.936
			1.0	0.981	0.101	0.196	2.965
			1.5	0.937	-0.211	0.167	1.558
			2.0	0.924	-0.355	0.940	1.005
500,000	100	20	0.0	1.001	-0.057	-0.054	4.730
			0.5	0.999	0.014	0.024	4.669
			1.0	0.993	0.185	0.229	3.223
			1.5	0.940	-0.235	0.122	1.420
			2.0	0.907	-1.054	0.530	1.009
500	400	80	0.0	0.214	0.000	0.520	0.002
			0.5	0.237	-0.001	0.523	0.002
			1.0	0.320	0.000	0.544	0.006
			1.5	0.530	-0.001	0.616	0.066
			2.0	0.733	-0.012	1.091	0.452
5,000	400	80	0.0	0.919	-0.003	0.061	2.687
			0.5	0.920	-0.001	0.064	2.505
			1.0	0.922	0.003	0.077	2.058
			1.5	0.930	-0.028	0.077	1.372
			2.0	0.940	-0.089	0.184	1.088
50,000	400	80	0.0	0.991	-0.008	-0.001	4.550
			0.5	0.991	0.004	0.012	5.276
			1.0	0.991	0.000	0.009	4.163
			1.5	0.980	-0.024	-0.001	1.777
			2.0	0.967	-0.171	-0.040	1.099
500,000	400	80	0.0	1.000	0.009	0.009	5.501
			0.5	0.999	0.001	0.001	5.180
			1.0	0.993	-0.001	0.006	3.852
			1.5	0.992	-0.022	-0.018	1.809
			2.0	0.971	-0.179	-0.079	1.136

We also considered the Sitter (1997) regression estimator, and repeated the entire simulation study that was performed using the ratio estimator in (4.1). Specifically, rather than (4.1), we made use of the estimator

$$\hat{\bar{Y}}_S^c = \bar{y} + b^*(\bar{x}' - \bar{x}), \quad (4.4)$$

which has approximate variance

$$V(\hat{\bar{Y}}_S^c) = (n^{-1} - m^{-1})S_d^2 + (m^{-1} - N^{-1})S_y^2, \quad (4.5)$$

where

$$S_d^2 = (N-1)^{-1}\sum_{i=1}^{N}[(Y_i - \bar{Y}) - \beta_{\text{POP}}(X_i - \bar{X})]^2$$

with

$$\beta_{\text{POP}} = \sum_{i=1}^{N} X_i Y_i / \sum_{i=1}^{N} X_i^2.$$

For each different combination of N, g, m, and n used in the simulation study, we computed

$$\hat{V}[(\hat{\bar{Y}}_S^c(t|k)] = (n^{-1} - m^{-1})s_{d(t|k)}^2 + (m^{-1} - N^{-1})s_{y(t|k)}^2, \quad (4.6)$$

for the t^{th} second phase sample drawn from the k^{th} first phase sample, where the sample variance

$$s_{d(t|k)}^2 = (n-1)^{-1}\sum_{i=1}^{n}[(y_{i(t|k)} - \bar{y}_{(t|k)}) - b_{(t|k)}^*(x_{i(t|k)} - \bar{x}_{(t|k)})]^2.$$

We also computed the jackknife estimator of variance

$$\hat{V}_{\text{JACK}}[(\hat{\bar{Y}}_S^c(t|k)] =$$

$$\frac{m-1}{m}\sum_{j=1}^{m}[\bar{y}_{(t|k)}(j) + b_{(t|k)}^*(j)\{\bar{x}'_{(t|k)}(j) - \bar{x}_{(t|k)}(j)\} - \{\bar{y} + b^*(\bar{x}' - \bar{x})\}]^2. \quad (4.7)$$

For each different combination of N, g, m, and n, equations (4.5) through (4.7) were used to compute values for RV, EBU, EBJ, and RE analogous to those given in Table 1 for the estimator in (4.1). The results obtained were extremely similar to those for the ratio estimator.

4.2 Simulation study: Chand (1975) and Ahmed (1997)

For purposes of the second set of simulations, we now assume that when the first-phase sample of m units is selected from the population of size N, information on two auxiliary variables X and Z is collected. When the second-phase sample of size n is selected from the first-phase sample, the study variable Y is measured, along with the two auxiliary variables X and Z. Note also that the auxiliary variable Z is assumed to be known for the entire population.

We began by creating a population of $N = 500$ units of (X_i, Z_i, Y_i) observations using

$$Y_i = \beta_1 X_i + \beta_2 Z_i + \varepsilon_i,$$

with $\beta_1 = 3.5$ and $\beta_2 = 2.5$. For each i, $i = 1, ..., N$, we generated X_i from a gamma distribution with a shape parameter of 2.2 and a scale parameter of one, Z_i from a gamma distribution with a shape parameter of 0.1 and a scale parameter of one, and ε_i from a standard normal. From the resulting population of (X_i, Z_i, Y_i) observations, we selected 1,000 first-phase sample of $m = 100$ units, and from each of these samples, we selected 10,000 second-phase samples of $n = 20$ units.

Following Chand (1975), a chain ratio estimator under two-phase sampling is given by

$$\hat{\bar{Y}}_{\text{Ch}}^c = \bar{y}(\bar{x}'/\bar{x})(\bar{Z}/\bar{z}'),$$

which has approximate variance

$$V(\hat{\bar{Y}}_{\text{Ch}}^c) = (n^{-1} - m^{-1})S_{d_2}^2 + (m^{-1} - N^{-1})S_{d_1}^2, \quad (4.8)$$

where

$$S_{d_2}^2 = (N-1)^{-1}\sum_{i=1}^{N}[(Y_i - \bar{Y}) - R_2(X_i - \bar{X})]^2$$

and

$$S_{d_1}^2 = (N-1)^{-1}\sum_{i=1}^{N}[(Y_i - \bar{Y}) - R_1(Z_i - \bar{Z})]^2$$

with

$$\bar{Y} = \sum_{i=1}^{N} Y_i/N,\ \bar{X} = \sum_{i=1}^{N} X_i/N,\ \bar{Z} = \sum_{i=1}^{N} Z_i/N,$$

$R_1 = \bar{Y}/\bar{Z}$, and $R_2 = \bar{Y}/\bar{X}$. In the simulation study, we computed

$$\hat{V}[(\hat{\bar{Y}}_{\text{Ch}}^c(t|k)] = (n^{-1} - m^{-1})s_{d_2(t|k)}^2 + (m^{-1} - N^{-1})s_{d_1(t|k)}^2, \quad (4.9)$$

for the t^{th} second phase sample drawn from the k^{th} first phase sample, where the sample variances

$$s_{d_2(t|k)}^2 = (n-1)^{-1}\sum_{i=1}^{n}[(y_{i(t|k)} - \bar{y}_{(t|k)}) - r_{2(t|k)}(x_{i(t|k)} - \bar{x}_{(t|k)})]^2$$

with

$$r_{2(t|k)} = \bar{y}_{(t|k)}/\bar{x}_{(t|k)}$$

and

$$s^2_{d_1(t|k)} = (n-1)^{-1}\sum_{i=1}^{n}[(y_{i(t|k)} - \bar{y}_{(t|k)}) - r_{1(t|k)}(z_{i(t|k)} - \bar{z}_{(t|k)})]^2$$

with $r_{1(t|k)} = \bar{y}_{(t|k)} / \bar{z}_{(t|k)}$. We also computed the jackknife estimator of variance

$$\hat{V}_{\text{JACK}}[(\hat{\bar{Y}}^c_{\text{Ch}}(t|k)] = \frac{m-1}{m}\sum_{j=1}^{m}\left[\bar{y}_{(t|k)}(j)\frac{\bar{x}'_{(t|k)}(j)}{\bar{x}_{(t|k)}(j)}\frac{\bar{Z}}{\bar{z}'_{(t|k)}(j)} - \bar{y}_{(t|k)}\frac{\bar{x}'_{(t|k)}}{\bar{x}_{(t|k)}}\frac{\bar{Z}}{\bar{z}'_{(t|k)}}\right]^2. \quad (4.10)$$

Using the same generated population of $N = 500$, we repeated the simulation; however we used $m = 400$ and $n = 80$ instead. We then created three additional populations of size $N = 500$ using $\beta_1 = 0.5$ with $\beta_2 = 0.5$, $\beta_1 = 3.5$ with $\beta_2 = 0.5$, and $\beta_1 = 0.5$ with $\beta_2 = 2.5$. For each of these three populations, we repeated the two simulations described above where in the first simulation, $m = 100$ with $n = 20$, and in the second simulation, $m = 400$ and $n = 80$. Finally, to study the effect of population size, we then repeated all the simulations based on the different values of β_1, β_2, m, and n when $N = 500$ for three additional values of N, namely 5,000, 50,000, and 500,000. For each different combination of N, β_1, β_2, m, and n, equations (4.8) through (4.10) were used to compute values for RV, EBU, EBJ, and RE analogous to those given in Table 1 for the estimator in (4.1). The results are provided in Table 2.

Generally speaking, the findings based on the results in Table 2 are similar to those arrived at for the estimators based on (4.1) and (4.4). In particular, the jackknife estimator of variance is more efficient than the usual estimator when the population size is sufficiently large. However, also of note is the fact that this efficiency seems to be related to the magnitude of the regression coefficients β_1 and β_2; that is, the jackknife estimator appears to achieve relatively greater efficiency for cases where the coefficient associated with the auxiliary variable X, is large relative to the analogous coefficient linked to Z.

Table 2
Comparison of the jackknife and usual estimators of variance of the chain ratio estimator of the population mean where the auxiliary variable, *X*, follows a gamma distribution with a shape parameter of 2.2 and a scale parameter of one, and the auxiliary variable, *Z*, follows a gamma distribution with a shape parameter of 0.1 and a scale parameter of one

m	*n*	β_1	β_2	*N*	RV	EBU	EBJ	RE
100	20	3.5	2.5	500	0.769	0.000	0.027	1.063
				5,000	0.831	-0.012	0.020	2.282
				50,000	0.818	-0.006	0.028	1.785
				500,000	0.852	0.001	0.036	1.993
100	20	0.5	0.5	500	0.911	-0.001	0.004	0.791
				5,000	0.943	-0.001	0.002	0.888
				50,000	0.948	0.000	0.003	0.896
				500,000	0.946	0.000	0.003	0.899
100	20	3.5	0.5	500	0.845	-0.001	0.015	1.674
				5,000	0.932	-0.011	0.000	3.632
				50,000	0.947	-0.005	0.004	3.221
				500,000	0.947	0.000	0.010	3.637
100	20	0.5	2.5	500	0.866	-0.001	0.009	0.668
				5,000	0.858	-0.003	0.008	0.775
				50,000	0.855	-0.001	0.010	0.670
				500,000	0.855	0.000	0.012	0.697
400	80	3.5	2.5	500	0.540	0.000	0.013	0.044
				5,000	0.780	-0.001	0.009	1.346
				50,000	0.819	0.000	0.008	1.878
				500,000	0.810	-0.001	0.006	1.953
400	80	0.5	0.5	500	0.817	0.000	0.003	0.254
				5,000	0.956	0.000	0.000	0.885
				50,000	0.973	0.000	0.001	0.946
				500,000	0.973	0.000	0.000	0.963
400	80	3.5	0.5	500	0.579	0.000	0.010	0.041
				5,000	0.907	-0.001	0.003	3.158
				50,000	0.954	0.000	0.002	3.845
				500,000	0.950	-0.001	0.001	4.853
400	80	0.5	2.5	500	0.787	0.000	0.004	0.222
				5,000	0.862	0.000	0.002	0.570
				50,000	0.873	0.000	0.003	0.698
				500,000	0.875	0.000	0.002	0.595

Finally, an analogous simulation study was performed using the regression estimator of Ahmed (1997). However, the populations were created using $\beta_1 = 10$ with $\beta_2 = 0.5$, $\beta_1 = 100$ with $\beta_2 = 0.5$, $\beta_1 = 0.5$ with $\beta_2 = 10$, and $\beta_1 = 10$ with $\beta_2 = 10$. As before when the estimators of Rao and Sitter (1995), Sitter (1997), and Chand (1975) were considered, provided that the population is sufficiently large, the jackknife estimator of variance seems to be more efficient than the usual estimator.

5. Conclusion and discussion

In this paper, the problem of estimating the variance of various estimators of the population mean in two-phase sampling has been considered by jackknifing the famous two-phase calibrated weights of Hidiroglou and Särndal (1995, 1998). Simulation studies based on ratio, regression, and chain-type estimators suggest that provided that the population size is large enough and the first and second-phase samples are relatively small, the jackknife estimator of variance is more efficient than the usual estimator of variance, regardless of the estimator for the population mean that is considered. For small populations, it might be worthwhile to adjust the finite population correction factors instead of directly applying the jackknife technique. This is an area where further research could be conducted.

Acknowledgements

This work was conducted while Sarjinder Singh was a postdoctoral fellow at Carleton University. The authors are grateful to the Associate Editor and the referees, whose comments greatly improved this manuscript. This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

Appendix

Derivation of the jackknife estimator in (2.1)

In this part of the appendix, we prove (2.1) for the jackknifed estimator of the population mean in two phase-sampling. First, note that $\hat{\beta}_1(j) = \hat{\beta}_1 + t_{1j}\, e_{1j}$ and $\hat{\beta}_2(j) = \hat{\beta}_2 + t_{2j} e_{2j}$, where $t_{1j} = q_{1j} w_{1j}^o z_j / (q_{1j} w_{1j}^o z_j^2 - \sum_{i \in s_1} q_{1i} w_{1i}^o z_i^2)$, $e_{1j} = x_j - \hat{\beta}_1 z_j$, $t_{2j} = q_{2j} w_{2j}^o x_j / (q_{2j} w_{2j}^o x_j^2 - \sum_{i \in s_1} q_{2i} w_{2i}^o x_i^2)$, and $e_{2j} = y_j - \hat{\beta}_2 x_j$. We also have $\hat{\bar{Z}}_1^o(j) = \hat{\bar{Z}}_1^o + h_{1j}(\hat{\bar{Z}}_1^o - z_j)$, $\hat{\bar{X}}_1^o(j) = \hat{\bar{X}}_1^o + h_{1j}(\hat{\bar{X}}_1^o - x_j)$, $\hat{\bar{X}}_2^o(j) = \hat{\bar{X}}_2^o + h_{2j}(\hat{\bar{X}}_2^o - x_j)$, and $\hat{\bar{Y}}_2^o(j) = \hat{\bar{Y}}_2^o + h_{2j}(\hat{\bar{Y}}_2^o - y_j)$, where $h_{1j} = w_{1j}^o / (1 - w_{1j}^o)$ and $h_{2j} = w_{2j}^o / (1 - w_{2j}^o)$.

Using these results, for $j \in s_2$, we have

$$\begin{aligned}\hat{\bar{Y}}^c(j) = {} & \hat{\bar{Y}}_2^o + \hat{\beta}_2(\hat{\bar{X}}_1^o - \hat{\bar{X}}_2^o) + \hat{\beta}_1\hat{\beta}_2(\bar{Z} - \hat{\bar{Z}}_1^o) \\ & + h_{2j}(\hat{\bar{Y}} - y_j) + t_{2j}\, e_{2j}(\hat{\bar{X}}_1^o - \hat{\bar{X}}_2^o) \\ & + \hat{\beta}_2\{h_{1j}(\hat{\bar{X}}_1^o - x_j) - h_{2j}(\hat{\bar{X}}_2^o - x_j)\} \\ & + t_{1j}\, e_{1j}\hat{\beta}_2(\bar{Z} - \hat{\bar{Z}}_1^o) - t_{1j}\, e_{1j}\hat{\beta}_2 h_{1j}(\hat{\bar{Z}}_1^o - z_j) \\ & + \hat{\beta}_1 t_{2j}\, e_{2j}(\bar{Z} - \hat{\bar{Z}}_1^o) - t_{2j}\, e_{2j}\hat{\beta}_1 h_{1j}(\hat{\bar{Z}}_1^o - z_j) \\ & - \hat{\beta}_1\hat{\beta}_2 h_{1j}(\hat{\bar{Z}}_1^o - z_j).\end{aligned}$$

Similarly, for $j \in (s_1 - s_2)$, we have

$$\begin{aligned}\hat{\bar{Y}}^c(j) = {} & \hat{\bar{Y}}_2^o + \hat{\beta}_2(\hat{\bar{X}}_1^o - \hat{\bar{X}}_2^o) + \hat{\beta}_1\hat{\beta}_2(\bar{Z} - \hat{\bar{Z}}_1^o) \\ & + \hat{\beta}_2 h_{1j}(\hat{\bar{X}}_1^o - x_j) + t_{1j}\, e_{1j}\hat{\beta}_2(\bar{Z} - \hat{\bar{Z}}_1^o) \\ & - t_{1j}\, e_{1j}\hat{\beta}_2 h_{1j}(\hat{\bar{Z}}_1^o - z_j) \\ & + \hat{\beta}_1\hat{\beta}_2\{(\bar{Z} - \hat{\bar{Z}}_1^o) - h_{1j}(\hat{\bar{Z}}_1^o - z_j)\}.\end{aligned}$$

Thus for $j \in s_2$,

$$\begin{aligned}\hat{\bar{Y}}^c(j) - \hat{\bar{Y}}^c = {} & \{\hat{\bar{Y}}_2^o(j) - \hat{\bar{Y}}_2^o\} - \hat{\beta}_2(j)\{\hat{\bar{X}}_2^o(j) - \hat{\bar{X}}_2^o\} \\ & - \hat{\beta}_1(j)\hat{\beta}_2(j)\{\hat{\bar{Z}}_1^o(j) - \bar{Z}\} \\ & + \hat{\beta}_2[\{\hat{\bar{X}}_1^o(j) - \hat{\bar{X}}_1^o\} - \hat{\beta}_1(j)\{\hat{\bar{Z}}_1^o(j) - \bar{Z}\}] \\ & + \hat{\beta}_2(j)\{\hat{\bar{X}}_1^o(j) - \hat{\bar{X}}_2^o\} \\ & + \hat{\beta}_2[\{\hat{\bar{X}}_2^o(j) - \hat{\bar{X}}_1^o(j)\} \\ & \quad - \hat{\beta}_1(j)\{\bar{Z} - \hat{\bar{Z}}_1^o(j)\} - \hat{\beta}_1\{\bar{Z} - \hat{\bar{Z}}_1^o\}],\end{aligned}$$

and for $j \in (s_1 - s_2)$,

$$\hat{\bar{Y}}^c(j) - \hat{\bar{Y}}^c = \hat{\beta}_2[\{\hat{\bar{X}}_1^o(j) - \hat{\bar{X}}_1^o\} - \hat{\beta}_1(j)\{\hat{\bar{Z}}_1^o(j) - \bar{Z}\}],$$

which proves (2.1).

Consistency of the estimator of variance in (2.2)

In this part of the appendix, we prove that the estimator $\hat{V}_{\text{JACK}}(\hat{\bar{Y}}^c)$ in (2.2) is consistent. First, note that the variance of the estimator $\hat{\bar{Y}}^c$ defined in (1.6) can be approximated as:

$$\begin{aligned}V(\hat{\bar{Y}}^c) \approx {} & V(\hat{\bar{Y}}_2^o) + \beta_2^2[V(\hat{\bar{X}}_1^o) + V(\hat{\bar{X}}_2^o) - 2\text{Cov}(\hat{\bar{X}}_1^o, \hat{\bar{X}}_2^o)] \\ & + \beta_1^2\beta_2^2 V(\hat{\bar{Z}}_1^o) \\ & + 2\beta_2[\text{Cov}(\hat{\bar{Y}}_2^o, \hat{\bar{X}}_1^o) - \text{Cov}(\hat{\bar{Y}}_2^o, \hat{\bar{X}}_2^o)] \\ & - 2\beta_1\beta_2\text{Cov}(\hat{\bar{Y}}_2^o, \hat{\bar{Z}}_1^o) \\ & - 2\beta_1\beta_2^2[\text{Cov}(\hat{\bar{X}}_1^o, \hat{\bar{Z}}_1^o) - \text{Cov}(\hat{\bar{X}}_2^o, \hat{\bar{Z}}_1^o)].\end{aligned}$$

If it is assumed that $\hat{\beta}_1(j) \approx \beta_1$, $\hat{\beta}_2(j) \approx \beta_2$, and similar to Rao and Sitter (1995), that $\bar{x}_n(j)/\bar{x}_r(j) \approx \bar{x}_n/\bar{x}_r$, it is quite straightforward to show that

$$\begin{aligned}\sum_{j\in s}[\hat{\bar{Y}}^c(j)-\hat{\bar{Y}}^c]^2 \approx{}& \sum_{j\in s_2}[\hat{\bar{Y}}_2^o(j)-\hat{\bar{Y}}_2^o]^2+\hat{\beta}_2^2\sum_{j\in s_2}[\hat{\bar{X}}_2^o(j)-\hat{\bar{X}}_2^o]^2\\ &+2\hat{\beta}_2\sum_{j\in s_2}[\hat{\bar{Y}}_2^o(j)-\hat{\bar{Y}}_2^o][\hat{\bar{X}}_1^o(j)-\hat{\bar{X}}_1^o]\\ &-2\hat{\beta}_2\sum_{j\in s_2}[\hat{\bar{Y}}_2^o(j)-\hat{\bar{Y}}_2^o][\hat{\bar{X}}_2^o(j)-\hat{\bar{X}}_2^o]\\ &-2\hat{\beta}_2^2\sum_{j\in s_2}[\hat{\bar{X}}_1^o(j)-\hat{\bar{X}}_1^o][\hat{\bar{X}}_2^o(j)-\hat{\bar{X}}_2^o]\\ &-2\hat{\beta}_1\hat{\beta}_2\sum_{j\in s_2}[\hat{\bar{Y}}_2^o(j)-\hat{\bar{Y}}_2^o][\hat{\bar{Z}}_1^o(j)-\hat{\bar{Z}}_1^o]\\ &-2\hat{\beta}_1\hat{\beta}_2^2\sum_{j\in s_2}[\hat{\bar{X}}_2^o(j)-\hat{\bar{X}}_2^o][\hat{\bar{Z}}_1^o(j)-\hat{\bar{Z}}_1^o]\\ &+\hat{\beta}_2^2\sum_{j\in s}[\hat{\bar{X}}_1^o(j)-\hat{\bar{X}}_1^o]^2\\ &+\hat{\beta}_1^2\sum_{j\in s}[\hat{\bar{Z}}_1^o(j)-\hat{\bar{Z}}_1^o]^2\\ &-2\hat{\beta}_1\hat{\beta}_2^2\sum_{j\in s}[\hat{\bar{X}}_1^o(j)-\hat{\bar{X}}_1^o][\hat{\bar{Z}}_1^o(j)-\hat{\bar{Z}}_1^o].\end{aligned}$$

Since the ten terms on the right hand side of this equation for $\sum_{j\in s}[\hat{\bar{Y}}^c(j)-\hat{\bar{Y}}^c]^2$ are the consistent estimators of the analogous ten terms in the equation above for $V(\hat{\bar{Y}}^c)$, it may be concluded that the jackknife estimator of variance in (2.2) is consistent.

References

Ahmed, M.S. (1997). The general class of chain estimators for the ratio of two means using double sampling. *Communincations in Statistics, Theory and Methods*, 26(9), 2247-2254.

Arnab, R., and Singh, S. (2006). A new method for estimating variance from data imputed with ratio method of imputation. *Statistics and Probability Letters*, 76, 513-519.

Berger, Y. (2007). A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika*, 94, 953-964.

Berger, Y., and Skinner, C. (2005). A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society*, Series B, 67, 79-89.

Chand, L. (1975). *Some ratio type estimators based on two or more auxiliary variables*. PhD Thesis, Iowa State Unversity, Ames, Iowa, USA.

Chen, J., and Shao, J. (2001). Jackknife variance estimation for nearest neighbour imputation. *Journal of the American Statistical Association*, 96, 260-269.

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

Fuller, W.A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica*, 8, 117-132.

Hidiroglou, M.A., and Särndal, C.-E. (1995). Use of auxiliary information for two-phase sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Vol. II, 873-878.

Hidiroglou, M.A., and Särndal, C.-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, 24, 11-20.

Kim, J.K., Navarro, A. and Fuller, W.A. (2000). Variance estimation for 2000 Census coverage estimates. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 515-520.

Kim, J.K., Navarro, A. and Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.

Kim, J.K., and Sitter, R.R. (2003). Efficient replication variance estimation for two-phase sampling. *Statistica Sinica*, 13, 641-653.

Kott, P.S., and Stukel, D. (1997). Can the jackknife be used with a two-phase sample? *Survey Methodology*, 23, 81-89.

Kovar, J., and Chen, E. (1994). Jackknife variance estimation of imputed survey data. *Survey Methodology*, 20, 45-52.

Raj, D. (1965). On sampling over two occasions with probability proportional to size. *Annals of Mathematical Statistics*, 36, 327-330.

Rao, J.N.K., and Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-60.

Singh, S. (2000). Estimation of variance of regression estimator in two phase sampling. *Calcutta Statistical Association Bulletin*, 50, 49-63.

Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.

Srivenkataramana, T., and Tracy, D.S. (1989). Two-phase sampling for selection with probability proportional to size in sample surveys. *Biometrika*, 76, 818-821.

Survey Methodology, June 2010
Vol. 36, No. 1, pp. 69-79
Statistics Canada, Catalogue No. 12-001-X

A comparison of sample set restriction procedures

Jason C. Legg and Cindy L. Yu [1]

Abstract

For many designs, there is a nonzero probability of selecting a sample that provides poor estimates for known quantities. Stratified random sampling reduces the set of such possible samples by fixing the sample size within each stratum. However, undesirable samples are still possible with stratification. Rejective sampling removes poor performing samples by only retaining a sample if specified functions of sample estimates are within a tolerance of known values. The resulting samples are often said to be balanced on the function of the variables used in the rejection procedure. We provide modifications to the rejection procedure of Fuller (2009a) that allow more flexibility on the rejection rules. Through simulation, we compare estimation properties of a rejective sampling procedure to those of cube sampling.

Key Words: Rejection sampling; Cube sampling; Stratification; Balanced sampling.

1. Introduction

A common practice in survey sampling is to utilize known population information about auxiliary variables to improve estimators of means and totals of characteristics of interest. When population control means or totals for an auxiliary variable are known, regression and other calibration estimators are often utilized. Let $(\boldsymbol{x}_i, y_i, p_i)$, $i = 1, 2, ..., N$, be a sequence of real vectors, where each $\boldsymbol{x}_i$ is a k dimensional vector, and a sample A be selected from $F_N = [(\boldsymbol{x}_1, y_1, p_1), ..., (\boldsymbol{x}_N, y_N, p_N)]$ using a sample design with inclusion probabilities p_i and joint inclusion probabilities p_{ij}. Suppose the population mean of $\boldsymbol{x}_i$, $\overline{\boldsymbol{x}}_N$, is known. Consider the regression estimator of the population mean of the form

$$\overline{y}_{\text{reg}} = \overline{\boldsymbol{z}}_N' \hat{\boldsymbol{\beta}}, \tag{1}$$

where $\boldsymbol{z}_i$ contains design variables and $\boldsymbol{x}_i$, $\overline{\boldsymbol{z}}_N$ is the population mean of $\boldsymbol{z}_i$, and $\hat{\boldsymbol{\beta}}$ is a regression coefficient estimator. For many designs, $\hat{\boldsymbol{\beta}}$ of the form

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i \in A} \boldsymbol{z}_i \, \phi_i \, p_i^{-2} \, \boldsymbol{z}_i' \right)^{-1} \sum_{i \in A} \boldsymbol{z}_i \, \phi_i \, p_i^{-2} \, y_i, \tag{2}$$

where ϕ_i are constants determined by the design, will be asymptotically efficient. Some examples of ϕ_i choices are $\phi_i = (1 - p_i)$ for Poisson sampling and for stratified random sampling, $\phi_{hi} = (N_h - 1)^{-1}(N_h - n_h)$ for element i in stratum h. If we assume there is a vector $\boldsymbol{d}$ such that

$$\phi_i \, p_i^{-2} \, \boldsymbol{z}_i' \boldsymbol{d} = p_i^{-1} \tag{3}$$

for all i, then estimator (1) is design consistent (Fuller 2002). The regression coefficient estimator (2) converges together with

$$\boldsymbol{\beta}_N = \left(\sum_{i=1}^{N} \boldsymbol{z}_i \, \phi_i \, p_i^{-1} \, \boldsymbol{z}_i' \right)^{-1} \sum_{i=1}^{N} \boldsymbol{z}_i \, \phi_i \, p_i^{-1} \, y_i.$$

As an example of applying equation (3), suppose we plan to select a Poisson sample and want to regress on a single covariate x_{1i} through the origin. If we add $(1 - p_i)^{-1} p_i$ into $\boldsymbol{z}_i$ to make $\boldsymbol{z}_i' = (x_{1i}, [1 - p_i]^{-1} p_i)$, then (1) will be design consistent for $\overline{y}_N$ since (3) is satisfied by setting $\boldsymbol{d}' = (0, 1)$. If we further assume that a column of ones is in the column space of the regression variables $\boldsymbol{z}_i$, then for these ϕ_i values, estimator (1) nearly attains the minimum asymptotic variance for design consistent regression estimators under certain regularity conditions (Rao 1994). An alternative approach to constructing a regression estimator is to start with a design consistent estimator, such as the generalized regression estimator of Särndal (1980), and determine the best coefficient given that form of the estimator. Starting with a design consistent form removes the need to satisfy (3). Condition (3) allows estimator (1) to be expressed in the form of a generalized regression estimator (Fuller 2009b, pages 116-117).

When auxiliary information is known at the unit level, the auxiliary information can also be incorporated into the sample design. For example in one classic case, the model with

$$y_i = \beta_0 + \beta_1 x_i + x_i \, \varepsilon_i, \tag{4}$$

$\varepsilon_i \sim \text{ind}(0, \sigma^2)$ and $\text{cov}(\varepsilon_i, x_i) = 0$ is assumed for the population F_N. From Isaki and Fuller (1982), the optimal inclusion probabilities for the regression estimator are those that are proportional to the square root of the design variances, *i.e.*, $p_i \propto x_i$ in this case. A possible sampling procedure is Poisson sampling with inclusion probabilities

1. Cindy L. Yu is an assistant professor in the Department of Statistics and the Center for Survey Statistics and Methodology at Iowa State University, Ames, IA 50010. E-mail: cindyyu@iastate.edu; Jason C. Legg is a postdoctoral researcher at the Center for Survey Statistics and Methodology at Iowa State University, Ames, IA 50010. E-mail: jason-legg@hotmail.com.

$$p_i = \left(\sum_{i=1}^{N} x_i \right)^{-1} n_N \, x_i, \tag{5}$$

where $n_N = \sum_{i=1}^{N} p_i$ is a specified target sample size. A second common design when model (4) is assumed is to stratify the population based on x. Strata are determined by setting the boundaries such that the sum of the sorted x_i values in each stratum are approximately equal. An equal number of units in each stratum are selected. This stratification design has the inclusion probabilities close to (5), and was shown to have an anticipated variance close to the best purposive sample model variance in the two-per-stratum case (Fuller 1981).

Another way to incorporate information from an auxiliary variable into the design is balancing. A sample A is balanced for variable $\boldsymbol{z}$ if

$$\overline{\boldsymbol{z}}_{\mathrm{HT}} = N^{-1} \sum_{i \in A} p_i^{-1} \boldsymbol{z}_i = N^{-1} \sum_{i=1}^{N} \boldsymbol{z}_i = \overline{\boldsymbol{z}}_N. \tag{6}$$

A design is balanced for $\boldsymbol{z}$ if every sample with positive probability is balanced for $\boldsymbol{z}$. Balancing can be thought of as calibration by design. To illustrate the effect of balancing, consider an equal inclusion probability design and $\boldsymbol{z}_i = (1, \, x_i)'$. The conditional prediction variance of $\overline{y}_{\mathrm{reg}}$ under model (4) is

$$V(\overline{y}_{\mathrm{reg}} - \overline{y}_N \,|\, \boldsymbol{x}, \, \overline{x}_{\mathrm{HT}}) = E\{V(\overline{u}_{\mathrm{HT}} \,|\, F_N) \,|\, \boldsymbol{x}, \, \overline{x}_{\mathrm{HT}}\} + (\overline{x}_N - \overline{x}_{\mathrm{HT}})^2 V(\hat{\beta}_1 \,|\, \boldsymbol{x}, \, \overline{x}_{\mathrm{HT}}), \tag{7}$$

where $u_i = x_i \, \varepsilon_i$. For a balanced design, the second term in (7) is 0, which suggests we might improve the estimator by balancing on x. In practice, a combination of balancing and calibration will often outperform either technique used alone.

Balanced sample designs have some additional practical value. For many designs, there is a nonzero probability of selecting a sample that contains undesirable auxiliary variable values. For example, an undesirable sample could be a sample with insufficient sample allocation for domains or a sample with a large number of extreme values of auxiliary variables. Although stratified designs reduce the set of such possible samples by fixing the sample size within each stratum, undesirable samples could still be possible. For example, some stratified samples might have some negative weights from using regression estimators. Balancing can remove poor performing samples by only retaining samples with estimates close to known quantities and with only positive weights for regression estimators.

Balanced sampling was proposed by Royall and Cumberland (1981) as a way to reduce model bias from incorrectly specified polynomial superpopulation models. Valliant, Dorfman and Royall (2000) discuss the implications of balancing from a prediction approach to sampling. Deville and Tillé (2004) investigated methods of selecting balanced samples within the design-based framework described above. See also Tillé (2006 Chapter 8) for a detailed treatment of balancing. In practice, finding a perfectly balanced design may not be possible. Very tight balancing can lead to a design with some extreme joint inclusion probabilities, including zero inclusion probabilities. Therefore, partial balancing is done in practice.

In this paper, we compare design properties through simulation studies of two balancing procedures, the rejective sampling of Fuller (2009a) and the cube sampling of Tillé (2006). We also provide modifications to Fuller's rejective sampling procedure that allow for more flexibility in balancing. In Section 2, the rejective sampling and the cube sampling are described. Properties of the inclusion probabilities of the two balancing procedures are compared in Section 3. In Section 4, some simulation results using balanced samples are presented. In Section 5, we provide adjustments to the rejective procedure. Concluding remarks are made in Section 6.

2. Balanced sampling procedures

Rejection sampling involves discarding any sample that does not meet a specified balancing tolerance. Fuller (2009a) presents one condition for rejecting a sample and Royall and Herson (1973) give another. In Fuller's procedure with the balancing variable vector $\boldsymbol{z}$, a sample is selected under a specified initial design and retained if

$$(\overline{\boldsymbol{z}}_{\mathrm{HT}} - \overline{\boldsymbol{z}}_N)' \, [V(\overline{\boldsymbol{z}}_{\mathrm{HT}} \,|\, F_N)]^{-1} \, (\overline{\boldsymbol{z}}_{\mathrm{HT}} - \overline{\boldsymbol{z}}_N) < \gamma \tag{8}$$

for some constant $\gamma > 0$, where $\overline{\boldsymbol{z}}_{\mathrm{HT}}$ is the Horvitz-Thompson mean estimator for variable $\boldsymbol{z}$, F_N is the given finite population,

$$V(\overline{\boldsymbol{z}}_{\mathrm{HT}} \,|\, F_N) = N^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} (p_{ij} - p_i \, p_j) \, \boldsymbol{z}_i \, \boldsymbol{z}_i' \, p_i^{-1} p_j^{-1},$$

p_i is the inclusion probability for unit i and p_{ij} is the joint inclusion probability of unit i and unit j under the initial design. Otherwise, the sample is rejected, a new sample is selected under the initial design, and condition (8) is checked for the new sample. If the original design has a central limit theorem, the left side of (8) is asymptotically a χ^2 random variable with degrees of freedom equal to the number of auxiliary variables. An approximate rejection rate can be set using the quantiles of a χ^2 distribution for γ. Choice of a rejection rate will depend on objectives of each individual survey. Low rejection rates may not reduce the variance by a large amount, but provide sufficient comfort to a researcher that a very poor sample will not be selected. On the other hand, high rejection rates could provide large reductions in the variance, but the resulting samples could

have insufficient sample size to accommodate unplanned domain analysis. For example, if a researcher decides to conduct domain analysis on the tail of the distribution of a balancing variable, the joint inclusion probabilities could be small leading to few units in the domain for many samples.

The cube method was developed by Tillé and Deville and is described in Tillé (2006). The cube method attempts to select a balanced sample with predetermined first-order inclusion probabilities. If the first-order inclusion vector does not lead to a balanced design, an additional step of minimizing a cost constraint is used. Unlike the rejection procedure, higher order initial inclusion probabilities are not prespecified. The cost minimization step maintains the specified initial first-order inclusion probabilities.

As a way to understand the cube procedure, Tillé (2006) describes sampling geometrically. The set of all possible samples is defined to be the set of vectors for vertices of an N dimensional unit cube. For example, if $N = 3$, the vertex $(0, 1, 1)$ denotes a sample containing units two and three. Using the balancing equation (6) and desired p_i for $i = 1, ..., N$, a balancing plane is created. Any sample where the balancing plane intersects a vertex of the unit N dimensional cube is a balanced sample. The design is balanced if every point of intersection between the balancing plane and the unit cube is a vertex of the unit cube. The cube sampling procedure begins by selecting a vector on the balancing plane, then a random walk from the initial point to an edge of the unit cube is done. Tillé refers to the random walk step as the flight phase. If the edge point at the end of the random walk is a vertex of the unit cube, the sample is selected. Otherwise, a cost minimization procedure is used to convert the fractional components of the edge vector to integers. The integer components of the edge vector are not changed in the cost minimization step. Tillé refers to the cost minimization step as the landing phase. Rejection sampling with high rejection rates produces results similar to cube sampling.

Other procedures besides rejection and cube sampling can be used to obtain nearly balanced samples. For example, stratification with boundaries determined by the $\boldsymbol{x}$ variables can also introduce some balancing effects to samples (Fuller 1981). Deciding the number of variables to use in the rejection and cube sampling procedures is essentially the same process as deciding how many variables to include in a regression estimator.

Software has been developed for selecting cube samples. For rejection sampling, standard software packages can be used to select a sample and compute (8). A loop needs to be written to complete the procedure. Programs for selecting cube samples have been written for SAS and R. See Rousseau and Tardieu (2004) for SAS and Matei and Tillé (2005) for R, and details of the procedures implemented are addressed in Deville and Tillé (2004). The R program available in the *sampling* library was used in the simulations in this paper. Because the cost minimization step of cube sampling is computationally intensive for more than 20 balancing variables, a variable suppression step is recommended for the landing phase in the programs.

3. Inclusion probabilities

Let π_i be the first-order inclusion probability for unit i and π_{ij} be the joint inclusion probability for unit i and j under a balanced design. Both rejective and cube sampling require initial first-order inclusion probabilities as inputs. The first-order inclusion probabilities are different than the initial values for rejection sampling. For rejection sampling, units closer to the population mean will have a slightly higher inclusion probability than units far from the mean. Cube sampling maintains the first-order inclusion probabilities from the initial specification. That is, for cube sampling $\pi_i = p_i$. Although for rejection sampling $\pi_i \neq p_i$, in general, the estimators considered will still use p_i rather than π_i.

To illustrate differences between initial and final inclusion probabilities, samples of size 20 from a population of 100 units were simulated. The population of x-values was generated as random variables from a standard normal distribution. The rejection procedure used simple random sampling as the initial design and balanced on x. The cube sample procedure used a balancing vector of $\boldsymbol{z}_i = (p_i, x_i)'$, where $p_i = 20/100$ for all i. The inclusion of p_i in the balancing vector for cube sampling was to control the sample size so that the resulting design would be comparable to using an initial design of simple random sample design in the rejection sampling simulation. First-order inclusion probabilities were estimated using a Monte Carlo simulation of size 100,000 (Figure 1). The curve was obtained by nonparametric fitting. An approximate 90% rejection rate was used for the rejection sampling. From rejection sampling theory, first-order inclusion probabilities are approximately a quadratic function of the distance $x_i - \bar{x}_N$ for an equal probability initial sample design (Fuller 2009a). The plot suggests that all first-order inclusion probabilities are 0.2 for the cube sample design. As expected, Figure 1 indicates the cube method maintains the specified first-order inclusion probabilities, but the rejective does not. As a result, the Horvitz-Thompson estimator using the initial inclusion probabilities (p_i) and the rejective samples is biased.

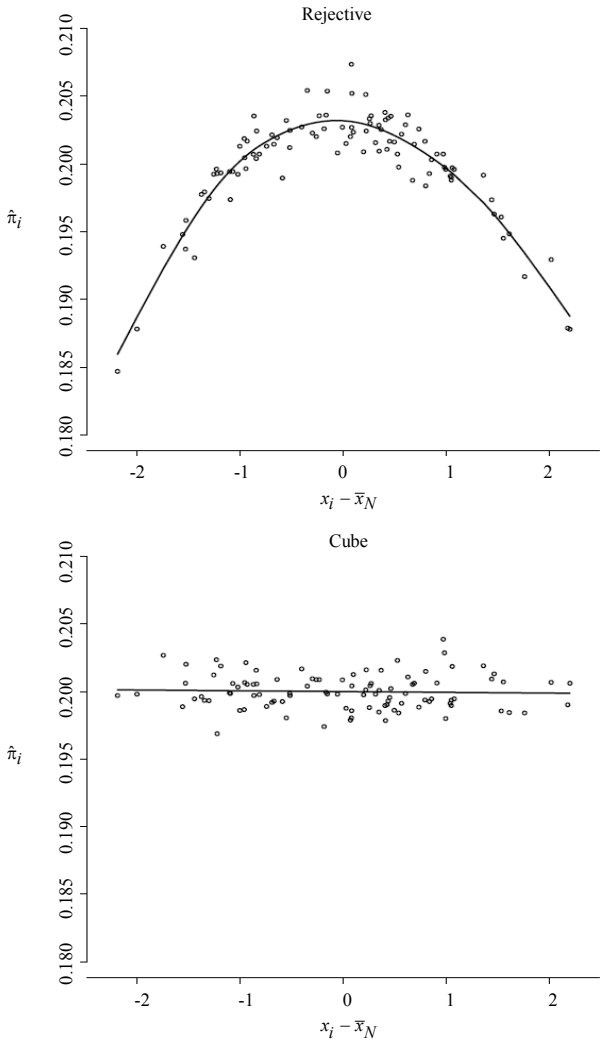

Figure 1 Simulated first-order inclusion probabilities. The balancing variable for the rejective method is $z_i = x_i$, and for the cube method is $z_i = (p_i, x_i)'$, where $p_i = 20/100$

The joint inclusion probabilities for the rejection sampling procedure differ from those of the initial design. A pair of units i and j are likely to have a high joint inclusion probability if $x_i + x_j - 2\bar{x}_N$ is close to zero for an equal probability initial sample design. The joint inclusion probabilities were estimated from simulated samples of size 20 from 100 (Figure 2). The joint inclusion probability for simple random sampling is 0.038. The rejection sampling joint inclusion probabilities are approximately a quadratic function of $x_i + x_j$. The plot of cube sampling joint inclusion probabilities against $x_i + x_j$ appears to have sharper angles than the rejection joint inclusion probabilities. High joint inclusion probabilities for the cube design are associated with pairs of units that are on the far opposite sides of $\bar{x}_N$. That is, for the sample value of $x_i + x_j$, those pairs with a large value of $|x_i| + |x_j|$ have a large probability of inclusion (Figure 3).

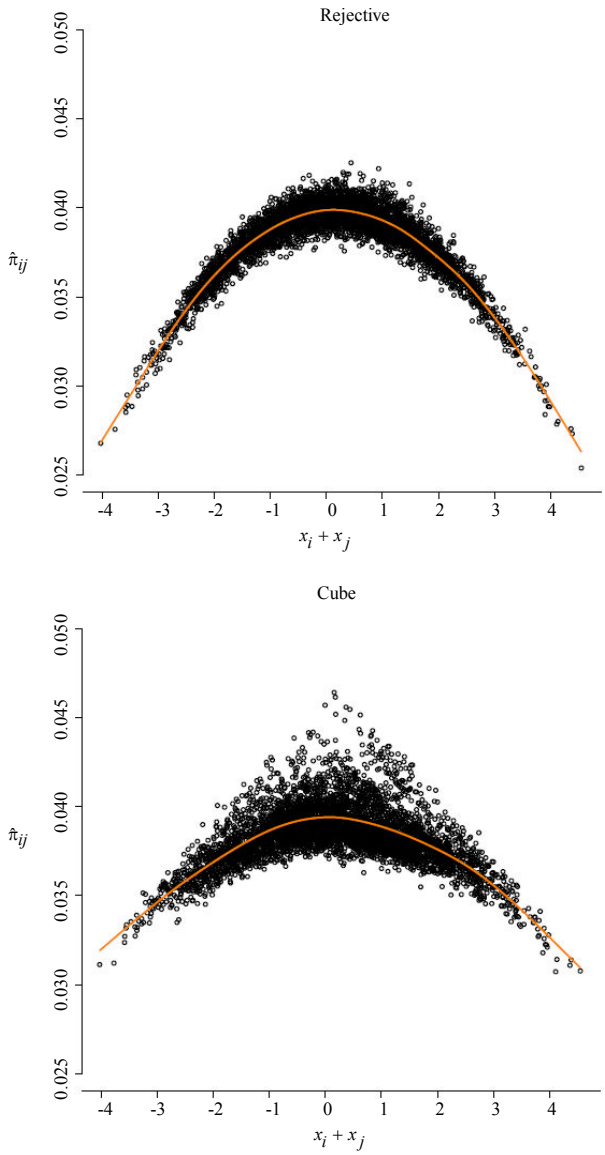

Figure 2 Simulated second-order inclusion probabilities. The balancing variable for the rejective method is $z_i = x_i$, and for the cube method is $z_i = (p_i, x_i)'$, where $p_i = 20/100$

The Horvitz-Thompson estimator using the initial inclusion probabilities under rejection sampling has an $O_p(n^{-1})$ bias while the Horvitz-Thompson estimator under cube sampling is unbiased. The standard Horvitz-Thompson variance estimator is biased for both procedures. Using Monte Carlo methods, the inclusion probabilities can be estimated so that nearly unbiased Horvitz-Thompson estimators can be used. However, for a large population, simulating enough samples to give a precise estimate of the joint inclusion probability for each pair of units is impractical. An alternative approach to variance estimation is to use a regression estimator and the variance estimator for the regression estimator. This is intuitively appealing because balancing is similar to regression through design.

Upon using the regression estimator, the bias of the regression estimator under both cube and rejective methods is of the same order. For rejective sampling, Fuller (2009a) gives conditions for the consistency of the variance estimator for the regression estimator. For cube sampling, Deville and Tillé (2005) and Tillé (2006) suggest using the variance estimator for a regression estimator furnishes a good approximation to the variance of the Horvitz-Thompson estimator. The variance estimators proposed by Deville and Tillé (2005) perform well when the joint inclusion probabilities of the resulting cube design are approximately equal to joint inclusion probabilities from a Poisson design. In the simulation studies of Section 4, the variance estimators proposed in Fuller (2009a) and Deville and Tillé (2005) are evaluated.

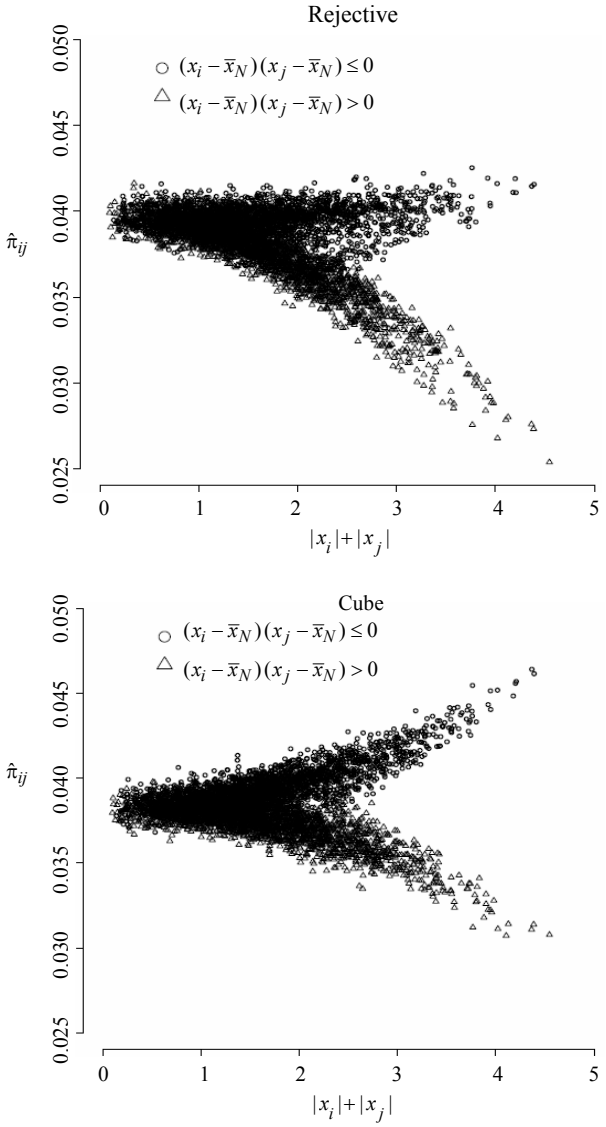

Figure 3 Simulated second-order inclusion probabilities with absolute sums of x. The balancing variable for the rejective method is $z_i = x_i$, and for the cube method is $z_i = (p_i, x_i)'$, where $p_i = 20/100$

4. Simulation of the regression estimator

A population of size 100 was generated from the model

$$y_i = x_i + 0.55\,x_i^2 + x_i\,\varepsilon_i \qquad (9)$$

$\varepsilon_i \sim$ iid $N(0, 0.4)$, where the x_i are fixed values in the range of 0 to 4 (Figure 4). Seventy-two of the x values were randomly simulated values less than 1.15 from a standard exponential distribution. The remaining 28 values, ranging from 0.18 to 4.0, were deterministically added to form the data set of x. The fixed x values were selected to be fairly right skewed so that some large and small strata when stratifying the population on x with approximately equal within-stratum sum of sorted x_i will be produced. The population was held fixed after initial selection. Model (9) contains a quadratic term, and was picked to simulate performance of the design and estimator strategy when model (4) was assumed in design and estimation.

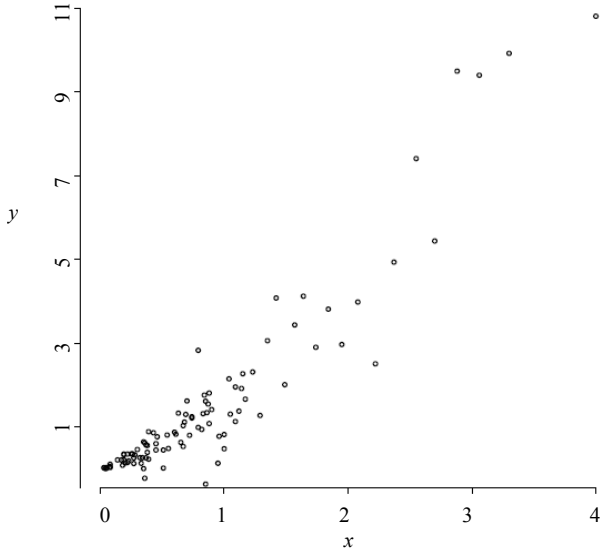

Figure 4 Simulation population under model (9)

We consider Poisson sampling and two-per-stratum stratified random sampling as initial designs. Strata were determined by setting the boundary so that the within stratum sum of sorted x_i was roughly equal for all strata. The sample size was set to 20, and ten strata were formed. The stratum sizes were 35, 15, 11, 9, 8, 7, 5, 4, 3, and 3. The rejection procedure used a stratified two-per stratum sample selection with equal inclusion probabilities within a stratum. The stratum boundaries were chosen this way so that the inclusion probability of unit i is closely proportional to x_i, which is the optimal inclusion probability under model (9) (Ikasi and Fuller 1982). Such a stratified design can also partially balanced on x through a standard design. Balance

in the stratified random sampling design is achieved using a step function to approximate a line. The stratified design will also be partially balanced on x^2. The stratified random sample design is intended to illustrate how much more one can benefit from additional balancing. Two units per stratum were drawn in order to obtain the maximum number of strata while still permitting unbiased variance estimation. Fuller (1981) showed that, in the two-per-stratum case, this stratified design has an anticipated variance close to the best purposive model variance under (4). Initial inclusion probabilities for the Poisson design with expected sample size 20 were set to the initial inclusion probabilities of the stratified design.

The regression estimator considered in this paper is in the form of (1) with $\hat{\boldsymbol{\beta}}$ defined in (2). The regression variable $\boldsymbol{z}$ is a vector of auxiliary variables that contains design variables and x. For the Poisson designs, we used $\boldsymbol{z}_i = (1,\ p_i,\ x_i,\ (1-p_i)^{-1} p_i)'$ as the vector of balancing variables and as the regression variable vector. The first variable provides control for population size, the second variable is a control for sample size, the third variable provides balance on x, and the fourth variable guarantees that the regression estimator is design consistent. See condition (3) for the design consistency of $\bar{y}_{\text{reg}}$ and set $\boldsymbol{d} = (0, 0, 0, 1)'$. For two-per-stratum stratified samples, the vector of balancing variables is $(x_i, I_{1i}, I_{2i}, ..., I_{10i})$ for cube sampling, where I_{hi} are the stratum indicator variables defined as

$$I_{hi} = \begin{cases} 1 & \text{unit } i \text{ in stratum } h \\ 0 & \text{otherwise} \end{cases}$$

for $h = 1, 2, ..., 10$. Only the x variable is included in the rejective balancing procedure since the sample from this initial design is automatically balanced on the stratum indicator variables. The regression variable vector for both balancing procedures is $\boldsymbol{z}_i = (x_i, I_{1i}, ..., I_{10i})'$.

For the initial designs, the variance estimators for $\bar{y}_{\text{reg}}$ are the variance estimators of the mean of $e_i = y_i - \boldsymbol{z}_i' \boldsymbol{\beta}_N$ calculated with $\hat{e}_i$, where $\hat{e}_i = y_i - \boldsymbol{z}_i' \hat{\boldsymbol{\beta}}$. For Poisson sampling, the variance estimator is

$$\hat{V}(\bar{y}_{\text{reg}}) = (n-s)^{-1} n \bar{\boldsymbol{z}}_N' \hat{\boldsymbol{M}}_{zz}^{-1} \sum_{i \in A} \boldsymbol{z}_i\, p_i^{-4} \times (1-p_i)^3\, \hat{e}_i^2\, \boldsymbol{z}_i' \hat{\boldsymbol{M}}_{zz}^{-1} \bar{\boldsymbol{z}}_N, \qquad (10)$$

where

$$\hat{\boldsymbol{M}}_{zz} = N^{-1} \sum_{i \in A} \boldsymbol{z}_i\, p_i^{-2} (1-p_i)\, \boldsymbol{z}_i',$$

and s is the number of variables in $\boldsymbol{z}$. Derivation of (10) is provided in the appendix.

For stratified random sampling with two-per-stratum, the variance estimator for $\bar{y}_{\text{reg}}$ is

$$\hat{V}(\bar{y}_{\text{reg}}) = (H-1)^{-1} H \sum_{h=1}^{H} [(1-W_h)^{1/2} \{0.5 W_h + (\bar{\boldsymbol{z}}_N - \bar{\boldsymbol{z}})\, \hat{\boldsymbol{M}}_{zz,h}^{-1}\, \phi_h\, W_h^2 (\boldsymbol{z}_{h1} - \boldsymbol{z}_{h2})\} \times (\hat{e}_{h1} - \hat{e}_{h2})]^2, \qquad (11)$$

where

$$\hat{\boldsymbol{M}}_{zz,h} = N_h^{-1} \sum_{i \in A_h} \boldsymbol{z}_i\, p_i^{-2}\, \phi_h\, \boldsymbol{z}_i',$$

A_h is the sample set in stratum h, $W_h = n_h / N_h$, $\phi_h = (N_h - 1)^{-1}(N_h - 2)$ for units in stratum h, $\boldsymbol{z}_{hi}$ is the auxiliary variable vector $\boldsymbol{z}_i$ in stratum h,

$$\hat{e}_{hi} = y_{hi} - \bar{y}_h - (\boldsymbol{z}_{hi} - \bar{\boldsymbol{z}}_h)' \hat{\boldsymbol{\beta}},$$

$\bar{y}_h$ and $\bar{\boldsymbol{z}}_h$ are stratum means of y_{hi} and $\boldsymbol{z}_{hi}$, respectively, and $H = 10$ is the number of strata. The derivation of (11) follows the same approach to the one in appendix and has been omitted.

For rejective sampling, the same variance estimators (10) and (11) using the initial design inclusion probabilities, were used to compute the variance estimator of $\bar{y}_{\text{reg}}$ for rejective samples. Fuller (2009a) proved that the large sample properties of the regression estimator for the rejective sample are the same as those of the regression estimator for the original inclusion procedure under some regularity conditions. For cube sampling, a variance estimator proposed by Deville and Tillé (2005) was evaluated for $\bar{y}_{\text{reg}}$ using cube samples.

Let $p(\cdot)$ denote the initial design and $\pi(\cdot)$ be the resulting scheme after balancing. The number of samples selected was 30,000 for each Monte Carlo simulation under initial designs, cube sampling and rejective sampling with both 90% and 95% rejection rates. The Horvitz-Thompson estimator $\bar{y}_{\text{HT}}$ and the regression estimator $\bar{y}_{\text{reg}}$ were constructed using initial inclusion probabilities p_i. Note that for rejection sampling, the Horvitz-Thompson estimator using the initial inclusion probabilities is not the Horvitz-Thompson estimator under the balanced designs. For each initial design, the following quantities were computed in the simulation studies.

- $V_p(\bar{y}_{\text{HT}})$ (or $V_p(\bar{y}_{\text{reg}})$): Monte Carlo variance of the Horvitz-Thompson estimator (or the regression estimator) using samples from initial designs.
- $V_\pi(\bar{y}_{\text{HT}})$ (or $V_\pi(\bar{y}_{\text{reg}})$): Monte Carlo variance of the Horvitz-Thompson estimator (or the regression estimator) for balanced samples.
- $\text{bias}_\pi(\bar{y}_{\text{HT}})$ (or $\text{bias}_\pi(\bar{y}_{\text{reg}})$): Monte Carlo bias of the Horvitz-Thompson estimator (or the regression estimator) using balanced samples.

For cube samples,

- $\hat{V}_{DT}(\bar{y}_{\text{reg}})$: estimated variance of the regression estimator using the variance estimators in Deville and Tillé (2005) and each cube sample.
- $\text{ave}(\hat{V}_{DT}(\bar{y}_{\text{reg}}))$: Monte Carlo average of $\hat{V}_{DT}(\bar{y}_{\text{reg}})$ using all cube samples.

Deville and Tillé (2005) recommend several variance estimators based on a Poisson sampling approximation with corrections for known constraints in the design variance. The first three estimators in Deville and Tillé (2005) have minor differences, therefore only the second estimator was used in the simulation studies. Deville and Tillé (2005) also propose the fourth estimator, but that estimator requires solving a nonlinear equation system, which would have been computationally expensive to add to the simulation. However, the fourth estimator could perform better than the other cases for stratified designs, since their fourth estimator reproduces the variance of a stratified random sample when the balancing vector contains stratum indicators.

For rejective samples,

- $\hat{V}(\bar{y}_{\text{reg}})$: estimated variance of the regression estimator using equation (10) (or (11)) for the Poisson (or two-per-stratum stratified) initial design and each balanced sample.
- $\text{ave}(\hat{V}(\bar{y}_{\text{reg}}))$: Monte Carlo average of $\hat{V}(\bar{y}_{\text{reg}})$ using all balanced samples.

In the simulations, $\hat{V}(\bar{y}_{\text{reg}})$ was also computed for cube samples, for comparison.

Table 1 reports the estimates for the Poisson design. The variance of the Horvitz-Thompson mean under initial Poisson sampling with expected sample size 20 and no balancing is $V_p(\bar{y}_{\text{HT}}) = 0.08$. The variances in Table 1 are standardized by $V_p(\bar{y}_{\text{HT}})$, and the biases are standardized by $\sqrt{V_p(\bar{y}_{\text{HT}})}$. The Horvitz-Thompson estimator is unbiased under the cube method designs, because cube sampling retains the first order inclusion probabilities. The Horvitz-Thompson estimator using initial design inclusion probabilities is biased under rejective sampling since the inclusion probabilities differ from the initial design inclusion probabilities, as indicated in Figure 1. The bias of the regression estimator under rejective sampling is less than the bias of the Horvitz-Thompson estimator with initial design inclusion probabilities. The bias of $\bar{y}_{\text{reg}}$ under both cube and rejective procedures is of the same order. Increasing the rejection rate increases the bias of $\bar{y}_{\text{reg}}$ for the rejection designs. However, the biases in $\bar{y}_{\text{reg}}$ under both balancing procedures and rejection rates are negligible relative to the Monte Carlo variances. For the Horvitz-Thompson estimator using initial design inclusion probabilities, the gain from using the balanced sample is substantial for both cube and rejective methods. The mean squared errors are further reduced by using the regression estimator along with either balancing procedures. The gain from using the regression estimator is larger for rejective sampling than for cube sampling, likely due to the cube method achieving tighter balance than the rejective method. Both procedures lead to similar variances for the regression estimator. The variance of the regression estimator under the Poisson initial design is $V_p(\bar{y}_{\text{reg}}) = 0.249$ (relative to $V_p(\bar{y}_{\text{HT}})$). By comparing 0.249 to the fourth row of Table 1, we can see that the gain from using the balanced samples on the regression estimator is moderate. The result is consistent with the finding in Fuller (2009a) that the variance reduction in $\bar{y}_{\text{reg}}$ by using rejective samples is due to a second order correction. The variance estimator of $\bar{y}_{\text{reg}}$ using (10) has small bias for both cube and rejective samples ($\text{ave}(\hat{V}(\bar{y}_{\text{reg}}))$ in Table 1). The variance estimator $\hat{V}_{DT}(\bar{y}_{\text{reg}})$ proposed in Deville and Tillé (2005) performed similarly as $\hat{V}(\bar{y}_{\text{reg}})$ in (10) since the second variance estimator in Deville and Tillé (2005) is very close to (10) for Poisson sampling. This result supports the claim that the Poisson approximation assumption in the variance estimators of Deville and Tillé (2005) is satisfied for the Poisson design case.

Table 1
Properties of samples based on Poisson sampling of expected size 20. $V_p(\bar{y}_{\text{HT}}) = 0.08$ and $V_p(\bar{y}_{\text{reg}})/V_p(\bar{y}_{\text{HT}}) = 0.249$

	Cube	Rej. 90%	Rej. 95%
$\text{bias}_\pi(\bar{y}_{\text{HT}})/\sqrt{V_p(\bar{y}_{\text{HT}})}$	-0.002	-0.016	-0.007
$\text{bias}_\pi(\bar{y}_{\text{reg}})/\sqrt{V_p(\bar{y}_{\text{HT}})}$	-0.002	0.002	0.005
$V_\pi(\bar{y}_{\text{HT}})/V_p(\bar{y}_{\text{HT}})$	0.142	0.270	0.220
$V_\pi(\bar{y}_{\text{reg}})/V_p(\bar{y}_{\text{HT}})$	0.131	0.136	0.129
$\text{ave}(\hat{V}(\bar{y}_{\text{reg}}))/V_p(\bar{y}_{\text{HT}})$	0.122	0.123	0.121
$\text{ave}(\hat{V}_{DT}(\bar{y}_{\text{reg}}))/V_p(\bar{y}_{\text{HT}})$	0.120	-	-

In Table 2, estimates under the initial two-per-stratum stratification design are reported. The variance of the Horvitz-Thompson mean under the initial stratification design is $V_p(\bar{y}_{\text{HT}}) = 0.011$ and all estimates are standardized by this value. Since stratification in this initial design controls for most of the effect of x on y, the regression estimator is not a large improvement over the Horvitz-Thompson estimator using initial design inclusion probabilities. The bias and variance of $\bar{y}_{\text{HT}}$ are close to those of $\bar{y}_{\text{reg}}$ under both cube and rejective methods. The larger estimated bias in $\bar{y}_{\text{HT}}$ under cube sampling is due to Monte Carlo error. The gain from balancing on x is not large, compared to the gain in the Poisson example. However, with this highly controlled initial stratified design, in which

the initial samples are already partially balanced on x, there still can be a modest benefit from additional balancing and using $\bar{y}_{\text{reg}}$ estimators. This result is seen for $\bar{y}_{\text{reg}}$ by comparing the fourth row of Table 2 to the variance of $\bar{y}_{\text{reg}}$ under the initial design $V_p(\bar{y}_{\text{reg}}) = 0.987$. Therefore, in this case a good strategy is to combine stratification, balancing, and regression, which is a similar conclusion drawn in Deville and Tillé (2004). The variance estimator $\hat{V}(\bar{y}_{\text{reg}})$ using (11) gives estimates on average for the regression estimator variances under both cube and rejective procedures that are close to the true variances. However, the variance estimator $\hat{V}_{DT}(\bar{y}_{\text{reg}})$ proposed by Deville and Tillé (2005) performed poorly for cube sampling. A possible reason is that the Poisson sampling approximation in the second variance estimator of Deville and Tillé (2005) assumes joint inclusion probabilities that are far from the actual joint inclusion probabilities in the small strata. The joint inclusion probabilities in the small strata are closer to those of stratified random sampling than Poisson sampling. This issue might explain why $\hat{V}(\bar{y}_{\text{reg}})$ in (11) using the initial two-per-stratum inclusion probabilities is less biased than $\hat{V}_{DT}(\bar{y}_{\text{reg}})$ in this case.

Table 2
Properties of samples based on stratified sampling of size 20. $V_p(\bar{y}_{\text{HT}}) = 0.011$ and $V_p(\bar{y}_{\text{reg}})/V_p(\bar{y}_{\text{HT}}) = 0.987$

	Cube	Rej. 90%	Rej. 95%
$\text{bias}_\pi(\bar{y}_{\text{HT}})/\sqrt{V_p(\bar{y}_{\text{HT}})}$	-0.028	0.014	0.010
$\text{bias}_\pi(\bar{y}_{\text{reg}})/\sqrt{V_p(\bar{y}_{\text{HT}})}$	-0.013	0.014	0.010
$V_\pi(\bar{y}_{\text{HT}})/V_p(\bar{y}_{\text{HT}})$	0.910	0.866	0.813
$V_\pi(\bar{y}_{\text{reg}})/V_p(\bar{y}_{\text{HT}})$	0.929	0.865	0.813
$\text{ave}(\hat{V}(\bar{y}_{\text{reg}}))/V_p(\bar{y}_{\text{HT}})$	0.907	0.881	0.775
$\text{ave}(\hat{V}_{DT}(\bar{y}_{\text{reg}}))/V_p(\bar{y}_{\text{HT}})$	0.792	-	-

To assess large sample properties of the balancing procedures, the size of the Poisson simulation was quadrupled. The population was replicated four times and a sample of expected size 80 was selected. The Horvitz-Thompson variance of a mean under the Poisson design is $V_p(\bar{y}_{\text{HT}}) = 0.020$ and the regression estimator variance is $V_p(\bar{y}_{\text{reg}}) = 0.132$. The resulting relative variances and biases are close to the results for samples of size 20 (Table 3). The simulation results agree with the theoretical result of Fuller (2009a) that the regression estimator is an $O_p(n^{-1/2})$ estimator after rejection of the type used in this paper. Although it has not been proven here, regression estimator after cube sampling appears to possesses similar properties to the regression estimator using rejection sampling.

Table 3
Properties of samples based on Poisson sampling of expected size 80. $V_p(\bar{y}_{\text{HT}}) = 0.02$ and $V_p(\bar{y}_{\text{reg}})/V_p(\bar{y}_{\text{HT}}) = 0.132$

	Cube	Rej. 90%	Rej. 95%
$\text{bias}_\pi(\bar{y}_{\text{HT}})/\sqrt{V_p(\bar{y}_{\text{HT}})}$	0.002	-0.006	-0.007
$\text{bias}_\pi(\bar{y}_{\text{reg}})/\sqrt{V_p(\bar{y}_{\text{HT}})}$	0.002	0.000	-0.001
$V_\pi(\bar{y}_{\text{HT}})/V_p(\bar{y}_{\text{HT}})$	0.127	0.267	0.224
$V_\pi(\bar{y}_{\text{reg}})/V_p(\bar{y}_{\text{HT}})$	0.122	0.124	0.123
$\text{ave}(\hat{V}(\bar{y}_{\text{reg}}))/V_p(\bar{y}_{\text{HT}})$	0.121	0.121	0.121
$\text{ave}(\hat{V}_{DT}(\bar{y}_{\text{reg}}))/V_p(\bar{y}_{\text{HT}})$	0.121	-	-

5. Adjustments to the rejection procedure

Fuller's rejection sampling procedure treats all balancing variables with the same importance. For a large number of balancing variables, exact balance on all variables cannot be expected and the approximation could be poor for some important variables. Therefore, a practitioner may want to have tighter balance on a subset of the balancing variables. As an example, a researcher may want to use Poisson sampling for simplicity but also have some control on the random sample size. A random sample size can complicate study planning and is a large contributor to the variance of estimators. Balanced sampling can be used to reduce the variation in sample sizes by balancing on the variable p_i, which is the initial first-order inclusion probability. For Fuller's rejection procedure, the variance of the sample size increases when the number of balancing variables increases and the rejection rate is held constant. The rejection procedure can be altered so that the p_i balance is tighter than the balance for other variables.

One approach to increasing the balancing on a subset of variables is to change the rejection test function. The order of the approximation to the first and second-order inclusion probabilities in Fuller (2009a) remains the same when the variance matrix in the rejection quadratic form is replaced with a symmetric positive definite matrix of the same order.

To determine weights for weighted rejection sampling, it is convenient to transform the balancing variables so that $V(\bar{z}_{\text{HT}} \mid F_N)$ is a diagonal matrix. The weighted rejection sampling test statistic is

$$\sum_{q=1}^{m} c_q V(\bar{z}_{\text{HT},q} \mid F_N)^{-1}(\bar{z}_{\text{HT},q} - \bar{z}_{N,q})^2, \qquad (12)$$

where m is the number of balancing variables, z_q is the q^{th} balancing variable, and c_q are selected weights. The weight on the first variable $z_{1i} = p_i$ can be set large relative to the weights on other variables to reduce variation in sample size. The transformation is the Gramm-Schmidt transformation using the design variances under the initial design. Balancing

is done on the transformed variables, but the first variable is not transformed. The transformed variables have uncorrelated Horvitz-Thompson estimators. Balancing on the transformed variables will still balance the original variables since each transformed variable is a residual from a regression operation on preceeding variables.

Equation (12) can be paralleled to the penalty term of the distance function underlying ridge calibration. See Rao and Singh (1997), Beaumont and Bocci (2008), and Chambers (1996). Specifically, selection of the c_q weights is similar to the problem of selecting appropriate costs in ridge calibration. Thus, rejection sampling using (12) can be viewed as incorporating ridge calibration at the design stage.

A second way to produce tighter balance on a subset of variables is to do rejection separately for subsets. A test statistic is produced for each subset and a sample must be accepted by all of the tests to be accepted. In the Poisson case, one test statistic may reject if the sample size is not within a specified tolerance of the expected sample size. This second approach requires some additional assumptions beyond those in Fuller (2009a), but a similar argument can be used to justify the procedure.

To prove the convergence properties of the multiple test rejection procedure, it is convenient to consider two subsets of balancing variables and think of rejection being done sequentially on each subset. We call the two subset rejection procedure a two-step rejective sampling procedure. Suppose $\boldsymbol{z}_i' = (\boldsymbol{z}_{1i}', \boldsymbol{z}_{2i}')$ is the balancing vector and the original design is denoted as $p(\cdot)$. The procedure is as follows.

Step 1: Select a sample using $p(\cdot)$ and reject samples with the balancing condition (8) on the first subset $\boldsymbol{z}_1$,

$$Q_1 = (\overline{\boldsymbol{z}}_{\mathrm{HT},1} - \overline{\boldsymbol{z}}_{N,1})' V(\overline{\boldsymbol{z}}_{\mathrm{HT},1} | F_N)^{-1} (\overline{\boldsymbol{z}}_{\mathrm{HT},1} - \overline{\boldsymbol{z}}_{N,1}) < \gamma_1.$$

Step 2: Use the accepted sample from step 1 to check the balancing condition (8) on the second subset $\boldsymbol{z}_2$,

$$Q_2 = (\overline{\boldsymbol{z}}_{\mathrm{HT},2} - \overline{\boldsymbol{z}}_{N,2})' V(\overline{\boldsymbol{z}}_{\mathrm{HT},2} | F_N)^{-1} (\overline{\boldsymbol{z}}_{\mathrm{HT},2} - \overline{\boldsymbol{z}}_{N,2}) < \gamma_2.$$

Reject the sample if the condition is not satisfied and repeat Step 1.

In both weighted and two-step procedures, trial and error is likely needed to choose γ's in practice. In the weighted procedure, the quadratic form becomes a sum of multiples of χ^2 random variables, which makes selection of γ more difficult than in the unweighted case. We used moment matching approximations to select γ's that provide rejection rates close to desired, but then resorted to small simulations to determine the rejection rate as a function of γ. For the two-step procedure, we used a χ^2 approximation to select a γ_1 that gave approximately the desired rejection rate at the first step, and used second χ^2 approximation to select an initial γ_2 that gave approximately the desired rejection rate at the second step. The second parameter γ_2 was adjusted in order to achieve the target overall rejection rate. The choice of γ's in the two-step procedure is subjective because many combinations of γ_1 and γ_2 can produce the same overall rate. In practice, a practitioner likely will set a tight bound for the first variable subset and loose bounds on the remaining balancing variables.

The large sample mean and variance of the regression estimator under the two-step rejective sample are the same as those of the regression estimator for the original design. Also, the usual estimator of variance under the original design for the regression estimator is appropriate for the two-step rejective sample. The proof of this statement is an extension of the proof in Fuller (2009a) and can be provided upon request.

To examine some properties of the two procedures, the Monte Carlo simulations for the Poisson initial sample design were repeated with the variable p_i separated from the other three variables. The balancing vector was transformed so that the variance matrix of the Horvitz-Thompson total estimators was diagonal. For the weighting procedure, the weight on the p_i component of the quadratic form was set to 1.5, the weights on the other components were set to 1, and γ was set to 0.627. This weighting procedure restricted the samples to those with sample sizes ranging from 18 to 22. For the two-step procedure, any sample with a sample size outside of the range from 18 to 22 was rejected in the first step and then the quadratic form for the remaining three variables was checked using a γ of 0.63 for the second step. Given the good performance of the variance estimator $\hat{V}(\overline{y}_{\mathrm{reg}})$ in (10), Table 4 only contains its Monte Carlo averages values $\mathrm{ave}(\hat{V}(\overline{y}_{\mathrm{reg}}))$.

Table 4
Properties of rejection samples with adjustments based on Poisson sampling of expected size 20, and 95% rejection rate

	Weighted	Two-step
$\mathrm{bias}_\pi(\overline{y}_{\mathrm{HT}})/\sqrt{V_p(\overline{y}_{\mathrm{HT}})}$	-0.005	-0.014
$\mathrm{bias}_\pi(\overline{y}_{\mathrm{reg}})/\sqrt{V_p(\overline{y}_{\mathrm{HT}})}$	0.003	0.002
$V_\pi(\overline{y}_{\mathrm{HT}})/V_p(\overline{y}_{\mathrm{HT}})$	0.210	0.217
$V_\pi(\overline{y}_{\mathrm{reg}})/V_p(\overline{y}_{\mathrm{HT}})$	0.132	0.132
$\mathrm{ave}(\hat{V}(\overline{y}_{\mathrm{reg}}))/V_p(\overline{y}_{\mathrm{HT}})$	0.121	0.121
$V_\pi(n)$	1.237	1.902

Results for expected sample size of 20 and a rejection rate near 95% were similar for the two adjustment procedures (Table 4). The Horvitz-Thompson estimator for the weighted procedure performed slightly better than the Horvitz-Thompson estimator for the two-step procedure. A reason for this discrepancy is that the weighted procedure

had much less variation in sample sizes ($V_\pi(n)$ in the last row of Table 4). Additional simulations with larger expected sample sizes gave similar relative variances. The regression estimator performed at roughly the same efficiency for the two procedures. The Horvitz-Thompson estimators using the initial design inclusion probabilities for these adjustment procedure performed slightly better than the Horvitz-Thompson estimator for the rejection procedure that did not place additional control on the sample size.

6. Discussion

Rejection sampling and cube sampling produce roughly equally performing regression estimators. Balancing provides major gains when the initial design provides little control on the auxiliary values entering samples. A well stratified sample design provides many of the benefits of balancing on a continuous variable. However, further balancing after stratification can still yield small mean squared error gains for regression estimators. Additionally, balancing could be used to prevent negative weights produced by regression estimators (Fuller 2009a).

For the simulations, the rejection rate was fixed at 90% for the larger population. When the population and sample sizes are increased, the rejection rate can be increased while still maintaining a large set of possible samples. Additional simulations were carried out with rejection rates near 99%, but the results were not presented since the differences between the results with 95% and with 99% were very small and the bias of $\bar{y}_{\text{reg}}$ remained negligible. The marginal variance reduction due to balancing decreases as the balancing condition is tightened.

In some special cases, an investigator may want to balance tightly on some variables and weakly on others. Gains can be made by choosing different weights for different variables or by dividing the variables into separate test sets. The weighted and two-step rejection procedures performed comparably, so the decision between procedures will largely be based on the ease of implementation.

Acknowledgements

This work was supported by Cooperative Agreement No. 68-3A75-4-122 between the USDA Natural Resources Conservation Service and the Center for Survey Statistics and Methodology at Iowa State University. The authors thank Wayne A. Fuller, the associate editor and two anonymous referees for helpful comments that improve the paper.

Appendix

Start with

$$V(\bar{y}_{\text{reg}} \mid F_N) = V(\bar{y}_{\text{reg}} - \bar{y}_N \mid F_N).$$

Let

$$\bar{y}_N = \bar{\boldsymbol{z}}_N' \boldsymbol{\beta}_N$$

and note

$$y_i = \boldsymbol{z}_i' \boldsymbol{\beta}_N + e_{Ni}$$

$$\hat{\boldsymbol{\beta}} = \left[\sum_{i\in A} \boldsymbol{z}_i \phi_i p_i^{-2} \boldsymbol{z}_i'\right]^{-1} \sum_{i\in A} \boldsymbol{z}_i \phi_i p_i^{-2} (\boldsymbol{z}_i' \boldsymbol{\beta}_N + e_i)$$

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_N + \left[N^{-1}\sum_{i\in A} \boldsymbol{z}_i \phi_i p_i^{-2} \boldsymbol{z}_i'\right]^{-1} N^{-1}\sum_{i\in A} \boldsymbol{z}_i \phi_i p_i^{-2} e_i. \quad (13)$$

Under assumptions (design consistency standard assumptions)

$$N^{-1}\sum_{i\in A} \boldsymbol{z}_i \phi_i p_i^{-2} \boldsymbol{z}_i' = N^{-1}\sum_{i\in U} \boldsymbol{z}_i \phi_i p_i^{-1} \boldsymbol{z}_i' + O_p(n^{-1/2}).$$

Write

$$N^{-1}\sum_{i\in U} \boldsymbol{z}_i \phi_i p_i^{-1} \boldsymbol{z}_i' = \boldsymbol{M}_{zz,N}.$$

Use the same argument to expand the $N^{-1}\sum_{i\in A} \boldsymbol{z}_i \phi_i p_i^{-2} e_i$ term. Then the expansion of (13) is

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_N + \boldsymbol{M}_{zz,N}^{-1} N^{-1}\sum_{i\in A} \boldsymbol{z}_i \phi_i p_i^{-2} e_i + O_p(n^{-1}).$$

For construction of confidence intervals for $\bar{y}_N$ it is enough to consider the variance of the linearized term. Therefore consider in the notation of Särndal, Swensson, and Wretman (1992),

$$\text{AV}(\bar{y}_{\text{reg}}) = \bar{\boldsymbol{z}}_N' \boldsymbol{M}_{zz,N}^{-1} V(\bar{\boldsymbol{b}}_{\text{HT}} \mid F_N) \boldsymbol{M}_{zz,N}^{-1} \bar{\boldsymbol{z}}_N$$

where

$$\boldsymbol{b}_i = \boldsymbol{z}_i \phi_i p_i^{-1} e_i.$$

The variance of the HT estimator for the mean of b_i under Poisson sampling is

$$\sum_{i\in U} (1 - p_i) p_i^{-1} \boldsymbol{b}_i \boldsymbol{b}_i'.$$

Next apply that $\phi = 1 - p_i$ to obtain the asymptotic variance approximation to the linearized part of $\bar{y}_{\text{reg}}$

$$\text{AV}(\bar{y}_{\text{reg}}) = \bar{\boldsymbol{z}}_N' \boldsymbol{M}_{zz,N}^{-1} \sum_{i\in U} (1 - p_i)^3 p_i^{-3} \boldsymbol{z}_i e_i^2 \boldsymbol{z}_i' \boldsymbol{M}_{zz,N}^{-1} \bar{\boldsymbol{z}}_N.$$

The variance estimator is obtained by replacing the population totals with HT estimators under Poisson sampling and incorporating a degree of freedom correction to the front of $n/(n-s)$ due to the small sample size.

References

Beaumont, J.-F., and Bocci, C. (2008). Another look at ridge calibration. *Metron*, 66, 1, 5-20.

Chambers, R.L. (1996). Robust Case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.

Deville, J.-C., and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91, 893-912.

Deville, J.-C., and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 569-591.

Fuller, W.A. (1981). An empirical Study of the ratio estimator and estimators of its variance: Comment. *Journal of the American Statistical Association*, 76, 78-80.

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

Fuller, W.A. (2009a). Some design properties of a rejective sampling procedure. Forthcoming *Biometrika*.

Fuller, W.A. (2009b). *Sampling Statistics*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Ikasi, C.T., and Fuller, W.A. (1982). Survey design under a regression population model. *Journal of the American Statistical Association*, 77, 89-96.

Matei, A., and Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics*, 21, 543-570.

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

Rao, J.N.K., and Singh, A.C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, D.C., 57-65.

Rousseau, S., and Tardieu, F. (2004). La macro SAS CUBE d'échantillonnage équilibré, Documentation de l'utilisateur. Technical report, INSEE, Paris.

Royall, R.M., and Cumberland, W.G. (1981). The finite-population linear regression estimator and estimators of its variance – An empirical study. *Journal of the American Statistical Association*, 76, 924-930.

Royall, R.M., and Herson, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68, 880-889.

Särndal, C.-E. (1980). On π inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.

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

Tillé, Y. (2006). *Sampling Algorithms*. New York: Springer Science+ Business Media, Inc.

Tillé, Y., and Matei, A. (2005). The R package Sampling. *The Comprehensive R Archive Network*, *http://cran.r-project.org/*, *Manual of the Contributed Packages*.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference*: *A Prediction Approach*. New York: John Wiley & Sons, Inc.

Survey Methodology, June 2010
Vol. 36, No. 1, pp. 81-93
Statistics Canada, Catalogue No. 12-001-X

The multidimensional integral business survey response model

Mojca Bavdaž [1]

Abstract

Knowledge of the causes of measurement errors in business surveys is limited, even though such errors may compromise the accuracy of the micro data and economic indicators derived from them. This article, based on an empirical study with a focus from the business perspective, presents new research findings on the response process in business surveys. It proposes the Multidimensional Integral Business Survey Response (MIBSR) model as a tool for investigating the response process and explaining its outcomes, and as the foundation of any strategy dedicated to reducing and preventing measurement errors.

Key Words: Accuracy; Data collection; Economic statistics; Business survey; Measurement error.

1. Introduction

Measurement errors represent the gap between an ideal measurement and the obtained survey response (Groves, Fowler, Couper, Lepkowski, Singer and Tourangeau 2004). To efficiently prevent or reduce the occurrence of measurement errors, it is necessary to know how the process of responding to survey questions evolves and what influences its course. Because work to reduce errors in business surveys has traditionally focused on sampling, frame, and nonresponse errors and, to a lesser extent, on measurement errors (Willimack, Lyberg, Martin, Japec and Whitridge 2004), knowledge of measurement errors and the underlying causal mechanisms is still largely limited in business surveys. This article attempts to fill that gap.

Most studies that examine the causes of measurement errors in business surveys are a product of pretesting research. As a result, most such studies are hypothetical (*e.g.*, Morrison, Stettler and Anderson 2002) or tentative (*e.g.*, Phipps, Butani and Chun 1995) as opposed to being based on actual data collection (*e.g.*, Hak, Willimack and Anderson 2003). The abundance of pretesting results, which are usually bound to a particular survey, contrasts with the scarcity of quality assessment research (*e.g.*, Giesen and Hak 2005) and with the shortage of generalization and linkages to the response process. Many studies focus on a particular aspect of the response process. For instance, Ponikowski and Meily (1989) examined the availability of data that business surveys require; Ramirez (1996) investigated respondent selection in business surveys; Jenkins and Dillman (1997) considered the design of business questionnaires; O'Brien (2000) and Willimack (2007) explored the respondent's role in the establishment survey response; Greenia, Lane and Willimack (2001) concentrated on business perceptions of confidentiality and on the closely connected issue of data sharing among statistical organizations; and Willimack (2003) exposed comprehension issues. Recently, more attention has been dedicated to the development and testing of electronic business questionnaires (*e.g.*, Snijkers, Onat and Visschers 2007) and their editing (*e.g.*, Nichols, Murphy, Anderson, Willimack and Sigman 2005), while more frequent complaints about the costs that statistical reporting imposes on the business community have triggered research on the response burden (*e.g.*, Hedlin, Dale, Haraldsen and Jones 2005).

The first study to systematically address the entire response process in establishment surveys was a general model of the survey response process for factual information, which Edwards and Cantor (1991) presented. Biemer and Fecso (1995) combined the cognitive model of Edwards and Cantor's (1991) survey response with a statistical model that tried to quantify measurement errors by their sources. Another attempt to grasp the entire response process in business surveys was made in 1998-1999, when the U.S. Census Bureau conducted unstructured qualitative interviews on statistical reporting. The study served as a basis for two business survey response models: the hybrid response model for establishment surveys by Sudman, Willimack, Nichols and Mesenbourg (2000) and the complete model by Willimack and Nichols (2001). Most recently, Lorenc (2006) suggested examining the entire response process on the basis of the idea of socially distributed cognition and using an establishment as a unit of observation.

These models identify many essential aspects of the response process in business surveys and offer some concepts for them, but they treat many issues only partially. This was an incentive for a comprehensive study of the response process of a selected business survey making possible further development of the business survey response model. This article presents the Multidimensional Integral Business Survey Response (MIBSR) model and discusses its contributions.

1. Mojca Bavdaž, Faculty of Economics, University of Ljubljana, Slovenia. E-mail: mojca.bavdaz@ef.uni-lj.si.

2. Empirical study

The aim of the empirical study was to build a conceptual framework of the response process – a response model – by examining from start to finish the actual response process to a typical business survey in a real business environment. The qualitative research interview was the primary method of investigation. The method was implemented using various techniques (mainly retrospective probing and ethnographic interviewing but also thinking aloud), two modes (in person and by telephone), and different interviewees (people from the participating business, questionnaire administration experts from the statistical organization, and subject-matter experts). In some cases on-site observation and analyses of micro data complemented those techniques. Considering all the variables, a range of approaches had to be developed (for more details, see Bavdaž 2009). On-site visits were arranged around two consecutive deadlines for the questionnaire's completion in 2005. An attempt was made to contact all key people involved in the response process.

The selected survey – the Quarterly Survey on Trade – was a business survey conducted by the Statistical Office of the Republic of Slovenia on a sample of approximately 1,600 legal units performing trade activities. It had classic characteristics of business surveys: a recurring mandatory governmental mail survey. Its instrument was an eight-page paper questionnaire and instruction and classification booklets. The questionnaire consisted of an introductory text and four sections, one referring to the business as a whole and the other three each referring to one kind of trade activity (commission trade, wholesale, and retail). All sections asked for sales and employment data. In addition, there were questions on sales breakdowns, stock, activity codes, and size and number of stores. Nonresponding units received up to three reminders and, ultimately, a telephone call. The final response rates were generally high, greater than 90%. Major deviations and inconsistencies discovered during editing procedures also required telephone calls to businesses.

The final sample in this study consisted of 28 businesses required to complete the Quarterly Survey on Trade. Previous studies resulting in models of the response process applicable to business surveys were based on small samples as well: 24 establishments (Edwards and Cantor 1991), 30 large multiunit companies (Sudman *et al.* 2000; Willimack and Nichols 2001), and 7 schools (Lorenc 2006). This is consistent with exploratory interview studies, which tend to have small sample sizes of "around 15 ± 10" (see Kvale 1996, page 102). The selection of businesses aimed to cover the heterogeneity of response processes. Because business size can be defined as the single most important business characteristic that is assumed to influence or be related to the characteristics of the response process (*e.g.*, O'Brien 2000), businesses were selected from all size classes.

Several measures boosted the validity of the research design. The businesses were selected from different size classes, including some of the largest ones in trade but also some from nontrade primary business activity. A few businesses refused to cooperate, mainly because of the work overload. Nevertheless, caution is necessary when applying findings to nontrade and overworked businesses. The study included people with different roles in the response process. Substantial effort was made to obtain participation and organize visits during the time the respondents were completing the questionnaire or right afterward so as to minimize the loss of information from their memory. The short time lags that occurred in some cases did not seem to be so damaging for remembering a frequently repeated and well-documented process, given the advance announcement of the impending on-site visit. Interview questions directed respondents to report how they last filled out the questionnaire (*e.g.*, when the books closed that month, how much time they spent, who signed the form and how fast), and respondents generally supported their reports by data from paper and electronic documentation they used to fill out the questionnaire. All this helped distinguish their last engagement from the usual one.

The interview as the primary research method was in some cases combined with observation. The interviews were tape-recorded and transcribed. More repeating patterns emerged as the fieldwork progressed, though diminishing returns of each consecutive on-site visit were noted toward the end of the fieldwork. The findings from the on-site visits were compared with the observations of the survey staff and subject-matter experts, quantitative data (where available), and previously published research. Alternative explanations were considered. Last but not least, the selection of a typical business survey made the generalization to other business surveys more plausible. As Yin (2003) suggests, all steps in the research were carefully documented to establish a chain of evidence and ensure high reliability of findings.

3. The MIBSR model

3.1 Presentation of the model

One of the main study results is the Multidimensional Integral Business Survey Response (MIBSR) model, which integrates previous research findings and new findings from my empirical study. The MIBSR model explicitly distinguishes between processes occurring at the individual level and others taking place at the organizational level, which is the business level in this case (see Figure 1). The cognitive

processes of comprehension, retrieval, judgment, and response occurring at the individual level are taken from Tourangeau's (1984) response model. They reflect the mental processes of people involved in the survey response that relate to the actual answering of particular survey questions as compared to the processes that refer to the organization, information support, and authorization of such answering, which occur at the business level. Contrary to the typical situation of surveys of individuals, parts of the process, such as requesting data from another participant or retrieving data from business records, are visible through participants' physical actions. By using the survey level, the MIBSR model also allows for the possibility of conceptualizing the response process over several implementations of a survey or over several surveys (indicated by the arrows in Figure 1).

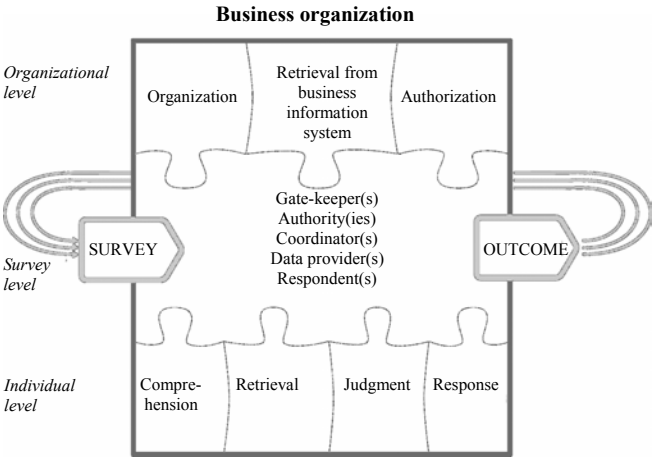

Figure 1 MIBSR model

The survey response task may involve several business participants who can enter and exit the response process at various points in time; but for the sake of clarity and simplicity, they are all depicted together. Business participants take part in organizational processes while going through their own cognitive processes; thus, they are a unifying link between processes at the individual and organizational levels. They may adopt one or more of the roles with a different influence on the response process, namely a gate-keeper (*e.g.*, a receptionist, boundary-spanning unit), an authority, a response coordinator, a data provider, or a respondent. Although Figure 1 presents participants from a single business organization, successful completion of the task may require either the participation of people who provide outsourced activities or communication with survey staff.

The response process is triggered when the survey instrument crosses the business's boundaries. The MIBSR model addresses the business response to a survey request presupposing a positive decision about participation in the survey. The examination of this decision, potentially leading to nonresponse, goes beyond the scope of this article even though it represents a natural introduction into the response process and may influence its course. The model suggests the most typical sequence of processes, although in practice some may be left out, repeated, or occurring in a different sequence. The following sections focus only on elaborated and newly added insights into the response process.

3.2 Organizational level

3.2.1 Organization of the survey response

Participation in a survey generally entails some preparatory activities due to work distribution and specialization in organizations. It requires an answer about who will perform the survey response task and when it will be done; both answers provide clues about how the task will be carried out. The study provided evidence that the two steps could be intrinsically linked. In fact, the selection of people for the survey response may itself indicate the priority assigned to the task in the organization. For instance, in some accounting firms and larger businesses, chiefs performed the task themselves, although they could have delegated the task, which may indicate a certain importance of the task, while the fact that many respondents received the task as novices may indicate its low priority. In contrast, priorities at the individual level were not always consistent with priorities at the organizational level. For instance, even if tax reporting gained higher priority than statistical reporting at the organizational level, this was irrelevant for a survey respondent not involved in tax reporting. I therefore examined the selection of business participants and the scheduling of the survey response task together within the organization of the survey response. The result is an expanded list of factors potentially influencing the organization of the survey response task (see Figure 2).

Tradition, customary practices, established procedures, and information location mainly influence the selection of business participants, which is an organizational matter, while other factors operate at both the organizational and the individual levels. Tradition dictates reliance on previous participants in recurring surveys when the same people repeatedly participate in the response process of the same (longitudinal) survey. Some study respondents claimed they had been "filling it out for years." Some had been filling it out since they started the job or since a colleague retired, went on a longer sick leave, left the job, and so on.

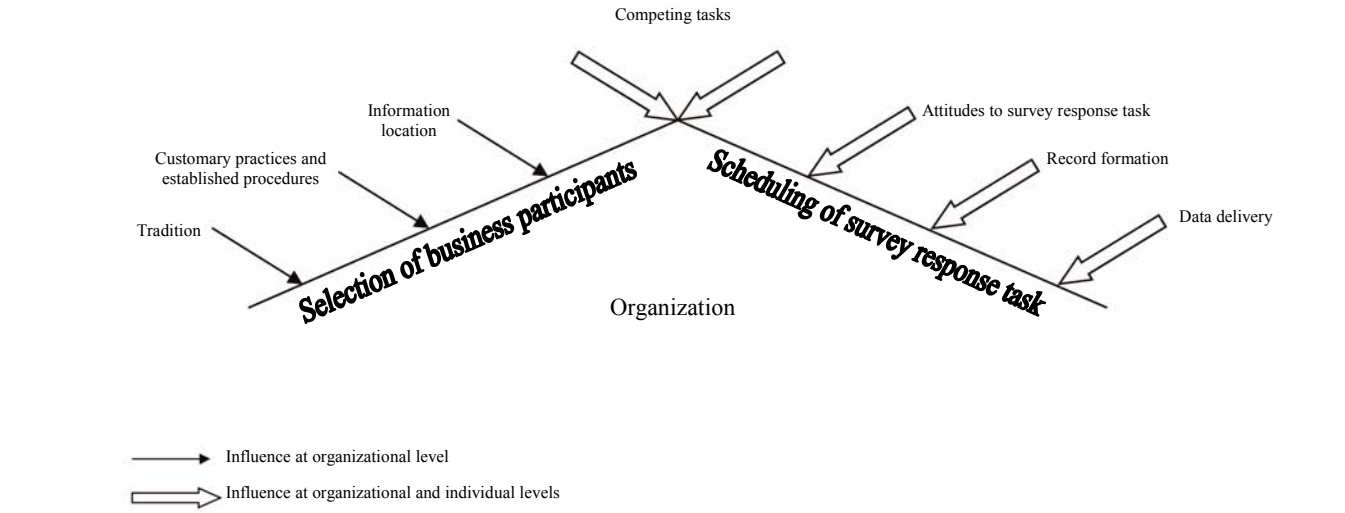

Figure 2 Factors influencing the organization of survey response

Many processes in organizations draw on customary practices and established procedures, which leads to the selection of the usual participants. This means that even when a new survey request reaches the business, the business will likely proceed in the same way as with previous survey requests because of the relatively stable distribution of work. In fact, some of the respondents in this study explained that the survey questionnaire would often be directed to the same department or person, who usually replied to such requests even if no formal policy on surveys existed. As one respondent clarified, "They prefer to bring them to me – this is the only policy." Some respondents knew which types of surveys they received, saying, for instance, "I'm doing all statistics except wages," or "I'm doing all statistics, also for the Bank of Slovenia, except Intrastat." Even in larger businesses, the same person often filled out several different survey questionnaires; one person completed all survey questionnaires that required financial data, be it for the Bank of Slovenia, the Statistical Office, or the Agency for Public Legal Records; others provided a list of specific surveys that they would complete, such as surveys on investments, fixed assets, value added, and so on.

Information location is an essential factor that influences the selection of business participants from the perspective of measurement errors. It refers to sufficient knowledge to provide an accurate survey response, including adequate access to records, if necessary. In this study, many respondents expressed that they had been chosen because of their access to data, for instance, "I have the data and I know how to retrieve them."

Competing tasks relates to the assignment of people and order to the tasks. It usually influences the choice of business participants at the organizational level when alternative possible participants are compared, as well as the scheduling of the survey response task at the individual level when the priorities of a participant's several tasks are considered. Study respondents in several, mainly smaller businesses agreed that they give low priority to the survey response task when they schedule their work: "VAT (value-added tax), debt recovery, bookkeeping . . . all has priority over statistics." Another respondent said that she "wouldn't think of doing the survey on the day all the book entries are done" but instead checks "the balance sheet, . . . liabilities, how the payments stand, how much debt there is, the financial situation." Another explained the work process as "internal reporting first, current affaires next, statistical reporting afterwards." In a few larger businesses, however, respondents said that they completed survey questionnaires as soon as data became available or final.

Similarly, attitudes to the survey response task can be examined at the organizational level through formal policies on surveys and the informal reactions of authorities as well as individual perceptions. Businesses in this study did not have any formal policies on surveys, though the discourse of authorities in some companies indicated their negative attitudes: "it's only statistics; prepare something." Organizational attitudes may affect the organization of the survey response, through potential consequences for the business, particularly opportunity costs, penalties, and damage to the public image. Most participants expressed a negative attitude toward surveys, describing them as "a necessary evil" and "redundant" or "additional" work. Individual attitudes toward surveys may contribute to the early, timely, or late scheduling of the task; they may also influence an individual's inclusion or exclusion in the survey response task.

Record formation and data delivery are primary in the scheduling of the response tasks. The timing of record formation determines when the records with required data about the business were created and took on the acceptable or desirable form, especially when the data become final. Respondents in larger businesses and businesses with foreign ownership typically referred to internal deadlines for "closing the books" or the VAT submission deadline. Data delivery is relevant in those cases where the participant must rely on other people to deliver required data. This particularly applied to accounting firms in this study. However, the timing of record formation and data delivery may vary by the kind of data requested, so that the latest record formation and the latest data delivery, eventually, determine the actual scheduling. For instance, some respondents explained that more time was necessary to get the correct value of stock because of lags in recording incoming invoices as compared to sales figures.

After the organization of the survey response task, the task can be realized, though it is sometimes necessary to further refine the selection of business participants or the scheduling to provide for all requested items, absence from work, and other circumstances.

3.2.2 Retrieval of information from the business information system

The capacity of the business information system (BIS) is the key factor that influences the response process and its outcome in business surveys. The BIS does not consist of the technological element only; it also includes people (Avison and Elliot 2006). The human capacity of the BIS relevant for the business survey response is mainly reflected in cognitive processes at the individual level (see section 3.3), while its technological capacity is determined through business records at the organizational level. The study showed that formation of business records depends on internal and external factors, though the line between the two groups is blurred (see Figure 3).

External factors – legal obligations, standards, and benchmark practices – are imposed on companies from the environment and dictate the content of business records through cogency or the threat of sanctions. Legislation, regulations, and other forms of power with the law set out legal obligations. With respect to that, study respondents mainly mentioned mandatory compliance with accounting standards and the requirements of tax authorities. The latter could refer to the business as a whole (*e.g.*, VAT reports) or to particular items (*e.g.*, excise duties on tobacco products). Other mandatory requirements may relate to contributions, securities, insurance, environmental issues, and so on. Participants usually noted the compulsory character of governmental business surveys, although the lack of sanctions for nonresponse or a late response made some participants question this; furthermore, changing record formation for statistical purposes only was unthinkable to most study participants. Standards are a softer form of external factors: they are not mandatory, but are expected to be followed in most cases. Two examples from the study include the use of a classification based on the European Article Number barcode standard and recommendations from accounting authorities. The study suggested that standards were not used in the case of specific reasons; for instance, the information systems of the smallest retailers did not support barcode use. Benchmark practices are the least influential group of external factors. They refer to good examples of practice that have gained some recognition and authority by reputation (and not by law or institutional power). For instance, some study respondents mentioned obsolete software versus current standards, while others stressed powerful capabilities of their software and its positive influence on data provision.

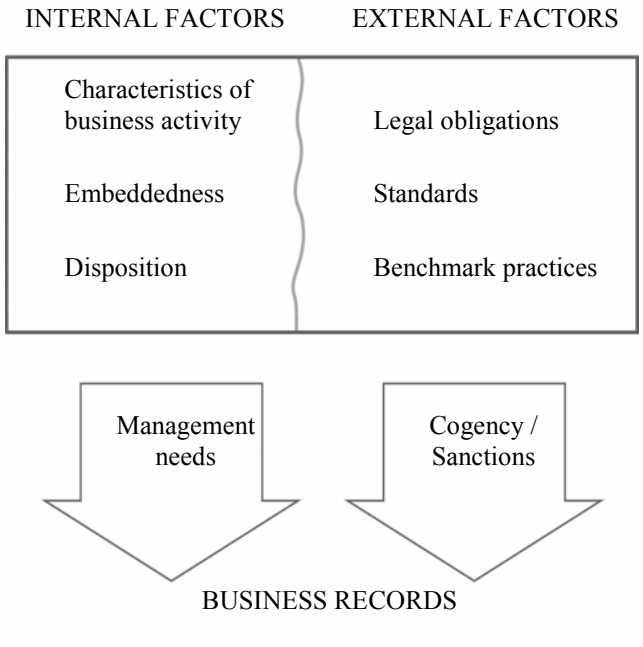

Figure 3 Factors of record formation

External factors drive data homogeneity and comparability in business records across companies, at least within similar economic activities. They provide the framework in which companies develop their own solutions for business records according to internal factors unless adhering to compulsory requirements more than fully satisfies data needs for running the business as was the case in small, local companies. Internal factors of record formation include characteristics of business activity, such as the size, type, and diversity of the business activity;

embeddedness in the business environment; and the disposition to forming records.

The size of the business activity plays a crucial role in record formation because it leads to a differential overview of an activity. In the study, most larger companies had an abundance of data. Business records provide information that cannot be gained from participation or observation only. That said, the size of the business activity is relative, especially if the size is observed only within legal boundaries or national borders. Therefore, it is better to speak about the embeddedness in networks of various kinds. In the study, for instance, a couple of smaller businesses had a foreign owner that demanded comprehensive reports to overcome the distance and manage the business remotely, and another small business had to use the sophisticated software of a business partner because it was its major supplier. The study also showed how different types of activities influenced the kind of available records; for example, wholesale businesses that typically put recipients on their invoices had more information on their buyers than businesses in retail that typically issued receipts without indicating the name. High diversity of business activities also is a major challenge for record formation in most businesses; in general, smaller businesses had renounced the use of detailed records and were forced to make estimates instead. Last, disposition refers to the prevailing attitudes of people in the business to various aspects of record formation, such as the inclination toward data, information technology, and change. Some businesses relied heavily on evidence-based decision making and thought highly of data; others showed enthusiasm for the possibilities of information technology, but a few others saw no usefulness in data.

Factors of record formation influence the availability of data in business records and their compliance with survey definitions. Data availability appears at the intersection of technological and human capacity in the business; knowledge is required to extract data from the BIS conditional on their existence. Several levels of answer availability in the BIS apply to survey questions (see Figure 4); their naming was inspired by the determination of cognitive states in Beatty and Herrmann (2002) and is in principle consistent with that proposed by Lorenc (2007):

(a) A datum is accessible – the required answer may be readily available. In this study, a typical example is total sales revenue, which is readily available to a person in accounting, or the number of employees, which is readily available to a person in the personnel department.
(b) A datum is generable – the required answer is not readily available to any person; the available data represent a basis for generating the required answer through manipulation. In the study, for instance, sales revenue in a particular trade activity was not always readily available, but it was possible to derive the exact figure by consulting two separate records (*e.g.*, the general ledger and commercial records).
(c) A datum is estimable – the required answer is not readily available to any person; the available data represent an approximation of the required answer or a basis for estimating the required answer through manipulation. In the study, a sales breakdown by commodity groups (*e.g.*, food, beverages, clothes, footwear) was often estimated by recategorizing available groups; however, those categories were sometimes too aggregated or too diverse to allow for an exact match (*e.g.*, Christmas products, Easter gifts, discontinued products).
(d) A datum is inconceivable – no available data lead to the required answer or its approximation; some bases for generating or estimating the required answer exist but require an unimaginable effort to produce it. For instance, a company would have to classify more than ten thousand invoices monthly to arrive at an exact breakdown of sales by kind of buyers.
(e) A datum is nonexistent – there are no bases for estimating the required answer. In the study, a cash-and-carry store could not distinguish between different kinds of buyers because they issued the same kind of nameless invoices to all customers, companies and individuals.

Because data availability varies across people in a business, it may be useful to determine answer availability at the individual level. In this case, a distinction has to be made between an answer that someone can obtain directly and an answer that they can access only through another person.

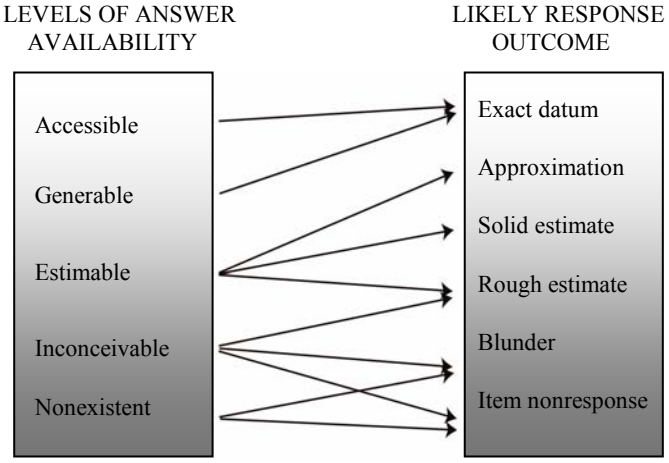

Figure 4 Levels of answer availability and likely response outcome

The final response outcome is conditional on the level of answer availability and may range from an exact datum to item nonresponse (see Figure 4). A measurement error occurs whenever the response outcome deviates from the exact datum. When a datum is accessible or generable, the response outcome is likely to be an exact datum, although the possibilities of committing a measurement error increase if data have to be accessed through other people or manipulated. When a datum is estimable, the response outcome may be an approximation with a negligible measurement error or an estimate with a minor or substantial measurement error. An inconceivable datum may, at best, lead to a rough estimate. When respondents have no adequate bases to provide a response, they may make wild guesses resulting in blunders or skip the question, which leads to item nonresponse.

3.2.3 Authorization of the business response

Authorization is the final opportunity for corrective actions before the business response is forwarded to the survey organization and documentation archived. Most businesses in this study found this organizational step inconsequential and even skipped it. In more than half of businesses, respondents signed the questionnaire themselves because "they have the mandate to sign such things" and "the director is very rarely present" or "does not deal with such things." Still, even in those cases, some respondents mentioned that the director had been informed about that procedure. In several businesses, the superior signed the questionnaire for the sake of formality and no verification procedures were in place because "the director trusts us" or "doesn't have the necessary data," or because "we work this way."

A superior was typically present in the largest companies, through formal authorization or informal notification. Internal verification was rare, which could be the consequence of preceding consultations with the superior. Accounting firms usually delivered the completed questionnaire to the business for signature, though businesses sometimes also signed the blank questionnaire in advance.

3.3 Individual level

Given the level of answer availability in the BIS, it rests on the performance of cognitive processes and accompanying physical actions (especially interaction with computers) at the individual level to determine the final response outcome. The MIBSR model proposes that three inherently linked types of knowledge are relevant for these processes: knowledge of business reality, knowledge of record formation, and knowledge of business records (see Figure 5). Although it may be difficult to disentangle the three types of knowledge in practice, the study seems to suggest that every type is particularly influential for one kind of cognitive process.

The division of cognitive processes into comprehension, retrieval, judgment, and response derives from Tourangeau's (1984) response model. In business surveys, these processes may not be defined as easily as in surveys of individuals because the initial organization of the response may involve only a brief and superficial consideration of the survey task with barely any impact on the later response process or a thoughtful reflection on the questions. The study mainly focused on respondents' cognitive processes because it is their task to answer survey questions. Nevertheless, observations of other business participants are provided where available.

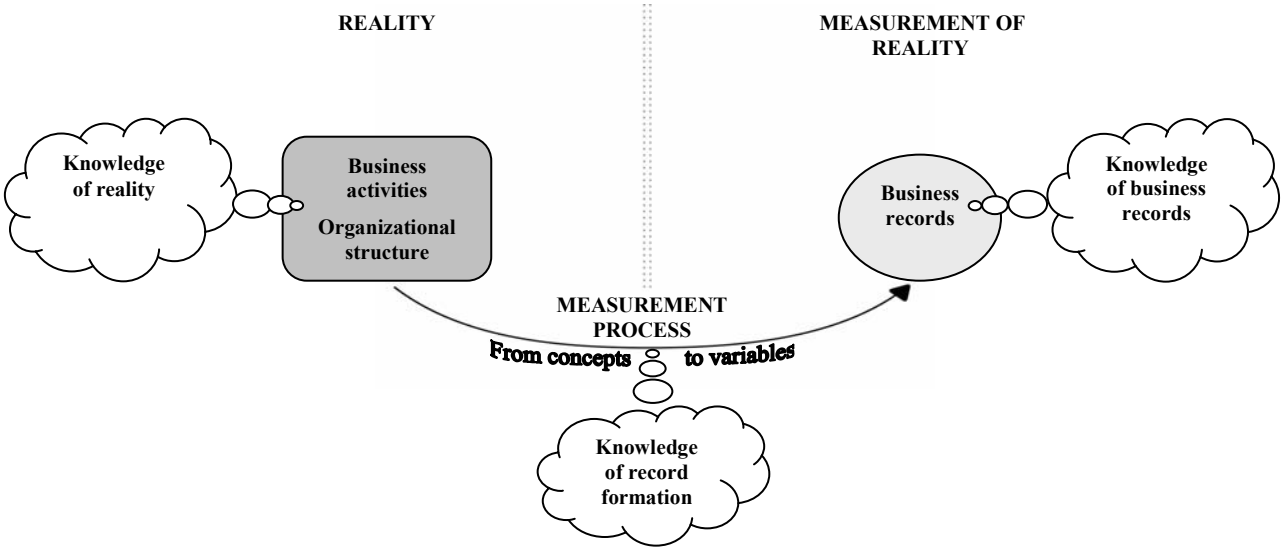

Figure 5 Knowledge relevant to the business survey response

3.3.1 Comprehension

In comprehension processes, respondents interpret the survey request for data, which usually is in the form of labels instead of questions. The MIBSR model suggests that, for comprehension processes, knowledge of business reality is particularly important. Business reality refers to the activities the business performs to subsist and to the division of work across locations and individuals. Knowledge of business reality thus presupposes acquaintance with every aspect of the business: who does what, what activities the business is involved in and how they are carried out, how decisions are made, why the business situation is as it is, how it evolved through time, and so on. Because larger businesses tend to be complex with technical and social divisions of labor, establishment of branches, organizational hierarchy, and decision-making structure (Tomaskovic-Devey, Leiter and Thompson 1994), it can be expected that fragmentation of the knowledge of the business's reality increases with business size.

This knowledge is essential in establishing whether survey questions are applicable to the business and providing correct answers afterward. In fact, no business in the study filled out all survey items. Respondents had to fill out only sections that applied to the kinds of trade they performed. Survey questions also required them to select applicable commodity groups, kinds of employment, kinds of buyers in wholesale, kinds of payment in retail, and so on. The required knowledge of business reality was occasionally specific: one respondent, for instance, needed information about the relationship between the company as the franchisor and their franchisees to avoid double counting or skipping some items across the businesses.

A major obstacle to using knowledge of business reality for correctly understanding survey questions was the incomprehension of economic and accounting concepts or their confounding with other concepts. For instance, one respondent had problems distinguishing between the concept of trade, which includes repackaging of goods, and the concept of production, which entails some transformation of goods beyond repackaging; a few respondents pondered over trade rendered on a commission basis because their activity was trade but accounting treated it as a service; many respondents associated retail with a store rather than with individuals as final consumers, regardless of the kind of buyer; one respondent defined wholesale as "everything that is not paid with cash" instead of linking it to nonfinal consumption; some respondents did not understand that "nontrade and nonmanufacturing organizations" were service providers; others did not understand the difference between merchandise and material, because the latter is an input to production (not trade) in accounting terminology and takes on another meaning in colloquial language, such as construction or building material.

Study respondents often used their own definitions to interpret survey questions. The same is true for those business participants who provided data on request without actually seeing the questionnaire and/or instruction booklet. This, for instance, happened in a few larger businesses where data providers completely relied on their own definitions of the sales space when providing data on store distribution by size of the sales space because additional explanations were given only in the instruction booklet.

3.3.2 Retrieval

In retrieval processes, the data and information required for the survey response are located and brought forth. In business surveys, the data usually reside in business records, not in memories, but knowledge is crucial for their extraction and interpretation. The retrieval thus mainly rests on knowledge of the business records, which refers to the contents and location of business records in the business and the possibilities of data access, including familiarity with applications and the people in charge of them.

Study respondents mainly exhibited good knowledge of the business records they worked with. In a couple of businesses where superiors participated in the response process, the superiors were not abreast of all details of the records and had an assistant perform the retrieval—but they had excellent insight into the business reality and knew how it converted into records. Even perfect knowledge of the business records, however, did not always suffice for exact answers. When the business records did not register all necessary data, knowledge of the business reality became critical for making correct inferences and good estimates. This sometimes happened in larger businesses and accounting firms where respondents knew the records very well, including the chart of accounts and its codes, but knew the assortment of merchandise only vaguely. As a result, they had to use estimates when classifying sales by commodity groups, as their acquaintance with the business activity was incomparable to a comprehensive, firsthand insight of sales personnel. In smaller businesses, lack of necessary data in records sometimes meant complete reliance on memory instead of records; a respondent, for instance, arrived at employment in wholesale by retrieving the number of people in relevant workplaces, namely chauffeurs, people who worked in the warehouse, salespeople, and office clerks.

3.3.3 Judgment

Judgment refers to the compilation of all retrieved data and information to formulate an answer. In this study, it frequently entailed some data manipulation or handling,

such as summation, balance with a residual, recategorization, and application of proportions. Judgment is mainly supported by knowledge of record formation. This knowledge provides information on how the business reality translates into business records and ensures that captured data are not considered isolated figures, codes, or words but take on a certain meaning representing the processes and objects measured. It therefore represents a link between knowledge of the business reality and knowledge of the business records (see Figure 5). Its importance was, for instance, noted during the observation of a respondent who was filling out the questionnaire and had to struggle with an inconsistency in the retrieved sales data. To identify the mistake, she systematically analyzed nonsales activities in the observed period and the correctness of their encoding in the records to finally discover a transaction that should not have been included in the sales figures.

However, lack of knowledge could not explain some judgments with an unfavorable response outcome, so the study looked more closely into principles that guided judgment. Among the most pervasive principles encountered in the survey response process under study was the principle of continuity, which advocates the use of the same response strategy in recurring surveys – even if this leads to errors. Continuity was sometimes considered within a year but also across years. It seemed to be strengthened by the lack of negative feedback from the statistical organization and its presumed satisfaction with the data. The study identified several respondents who used detailed procedures of calculation that were quite obsolete. A respondent even erroneously left out the section of commission trade but would not change the procedure during the year to avoid disrupting the reported data.

Two other principles were identified in relation to the principle of continuity: the principle of consistency and the principle of disregarding the exceptional. The principle of consistency implies use of the same or similar response strategies in the same survey questionnaire. For instance, a respondent who attributed various items of merchandise to only one commodity group in wholesale did the same in retail; a respondent who estimated wholesale turnover from VAT figures used the same approach to retail turnover, and so on. The principle of disregarding the exceptional implies ignoring new, one-off, or temporary activities. For instance, a study respondent inadvertently reported a temporary activity not reported in the questionnaire; another confessed the exclusion of new activities from reporting because their success was uncertain. The question, however, is how to set boundaries on the novelty and on the temporariness and when precisely such activities become representative of the business.

The principle of disregarding the exceptional is also related to the principle of disregarding the marginal, which advises ignoring those activities that are perceived as marginal to the business. For instance, some study respondents disregarded some items in sales breakdowns if they represented less than one percent of activity. The impact of the principle depends on the use of the collected data. It should be inconsequential if the aim is to estimate national totals or change. However, sales of a specific commodity group may be marginal to a large business but not marginal for the market of that commodity group.

The business perspective principle advocates the priority of the business perspective as compared to a statistical request. In the study, data on existing organizational units were judged acceptable despite their divergence from the required units; data on various packages (*e.g.*, a newspaper supplemented with a book) that were relevant from the business perspective were not disentangled for statistical purposes.

3.3.4 Response

The response component refers to the processes of mapping a judgment onto a response category and editing the response (Tourangeau, Rips and Rasinski 2000). In business surveys, mapping usually translates into matching available data from the BIS with response categories offered, which provides room for a specific form of measurement error: misclassification. For instance, when respondents had problems fitting available sales data into the provided classification scheme, they often chose the closest category, the main category, or the category "other."

The study also identified the presence of editing processes that show different aspects of business sensitivity. Some study respondents checked whether their selection of the decisive activity code was consistent with their registered activity, which may show a fear of nonconformity with administrative requirements. Not reporting people who helped in family businesses may reveal tax evasion. Although many respondents agreed that the data they reported in the questionnaire were considered confidential, there was scarce evidence of hindrance for disclosing the data to the statistical organization (*e.g.*, not reporting detailed data on newly introduced activities).

3.4 Survey level

The MIBSR model introduces the possibility of conceptualizing the response process over several implementations of a survey or over several surveys. It thus conceptually enables the observation of how the elements of survey design, which is under the control of survey organization, influence the response process.

The study focused on the impact of recurrence on the response process. In repeated administrations of the survey to the same business, the organization of the survey response became less relevant or irrelevant if it was a perfect replica of the preceding administration. The cognitive processes at the individual level were characterized by routine when the same business participants performed them. Many respondents admitted that they had not read the whole questionnaire, let alone the instructions in a repeat questionnaire. This also occurred in businesses that agreed to be observed while completing the questionnaire: after respondents gave the questionnaire a swift scan for any changes, they plunged into the retrieval processes based on the previously completed questionnaire or on other documentation and supporting notes. The comprehension step was thus performed superficially and pertained more to understanding completion of the previous questionnaire than it did to understanding survey requests. The retrieval procedures followed the previously established course and exhibited learning-curve effects. The respondent's judgment clung to the initial approach and was unlikely to change. The recurrence frequently loosened up a respondent's supervision and reduced the importance of the authorization or even omitted it.

Given the appointment to the survey task of the same people or usual units in the business, many of them sooner or later had contact with survey staff, despite the common self-administrative mode of data collection in business surveys. Such contact could occur early in the response process and influence the respondent's comprehension and judgment. This was rarely the case in the study; only a few respondents asked for explanations the first time they participated in the survey and another respondent asked for help when the business's activity changed. Contacts in which respondents requested postponement of the deadline did not seem to influence the subsequent response process, though the same could not be claimed for respondents who resisted participation. All other contacts happened during a follow-up when the response process, or parts thereof, had to be performed again, which could result in an adjusted survey response. Although respondents mainly acknowledged the politeness of the survey staff, their calls signaled that something was wrong: a missed deadline, an item missing in the questionnaire, an inconsistency in the reported data. The rareness of such contacts made a significant impression on respondents because these contacts were often the only type of feedback from the statistical organization.

In contrast, respondents did not always appreciate a lack of feedback. They expected feedback from the statistical organization after they first participated in the survey, but this generally did not happen. The lack of reaction made them confident in their approach, thus reinforcing the principle of continuity in their judgment. However, many respondents reported at least one piece of data that was not completely accurate (or not as accurate as they would expect the data should be) and they perceived the lack of complaints as satisfaction with bad data. Some respondents were convinced that the statistical organization knew about their business activity, which is why they rarely provided textual descriptions of seasonal oscillations. Given these observations, it is not surprising that several respondents expressed doubts about the accuracy of statistical data or questioned the accuracy of data that others provided. The right feedback may not only be important for that particular survey but also for participation in other surveys because it contributes to general perceptions on surveys and statistics.

4. Discussion of model's contributions

The dominance of written communication between the survey organization and businesses has moved business participants away from the center of statistical production and reduced the possibilities of insights into the process of responding to survey requests and the causes of measurement errors. By studying the response mechanisms and influencing factors, response models help bring these insights out and design approaches that turn this knowledge into an advantage. This section discusses the contributions of the MIBSR model with respect to previous response models applicable to business surveys.

4.1 Model construction

Two approaches were encountered in construction of previous models: adding some organizational steps to the core cognitive processes from Tourangeau's cognitive model of survey response (Biemer and Fecso 1995; Edwards and Cantor 1991; Sudman *et al.* 2000; Willimack and Nichols 2001) or using the organization as the unit of observation (Lorenc 2006). The MIBSR model explicitly links the processes to the level at which they occur: cognitive processes to the individual level and organizational processes to the organizational (in our case, the business) level. It also foresees the observation of the response process over several implementations of the same survey or over several surveys with different designs, which is particularly interesting for governmental surveys. By analyzing complex response processes at the appropriate level of observation, the MIBSR model sets up a framework that can also be used for quantitative modeling and experimental design.

4.2 Insights at the organizational level

Previous models treated initial organizational arrangements in the context of respondent selection (Biemer and Fecso 1995; Edwards and Cantor 1991) or in separate steps of respondent selection and the assessment of priorities, the latter ranking statistical reporting to the government lower than most other business reporting activities (Sudman *et al.* 2000; Willimack and Nichols 2001). They also identified several factors that influence respondent selection, especially the functional role, authority level, and position with regard to the information system (Edwards and Cantor 1991), knowledge of the information system, terms and definitions (Biemer and Fecso 1995), competing job responsibilities and access to the data (Sudman *et al.* 2000). The MIBSR model integrates all preparatory activities in the organization of survey response and suggests an expanded list of influencing factors. The organization of survey response now acknowledges that delegation of the task may also include selection of other business participants beyond respondents and that priority of competing tasks is just one of the factors influencing the task's scheduling.

All previous models have paid considerable attention to record formation. The MIBSR model suggests a different systematization and extension of factors of record formation, initially grouped into management, regulation, and standards by Willimack and Nichols (2001). Because it is generally unlikely that the requirements of statistical reporting are an actual factor of record formation, the MIBSR model may assist the survey organization in its endeavors to exert influence on record formation and eventually obtain requested data. Taking into account technological and human capacity of the BIS, the MIBSR model defines several levels of answer availability based on the extent to which the answer conforms to required survey definitions and proposes the likely response outcome. In authorization of the business response, the MIBSR model reiterates the possibility of internal verification that Sudman *et al.* (2000) and Willimack and Nichols (2001) propose for the release step. Authorization is more likely sought out when the survey response involves legally separate units and more formalized and centralized organizations.

4.3 Insights at the individual level

At the individual level, which deals with comprehension, retrieval, judgment, and response (Tourangeau 1984), the MIBSR model further elaborates on the knowledge relevant to cognitive processes. Willimack and Nichols (2001) emphasized personal knowledge for answers directly from memory and knowledge of the records. The MIBSR model suggests that a thorough understanding of the data in business records and their appropriate use in the survey response require knowledge of the whole chain of data generation, from knowledge of business reality to knowledge of record formation and knowledge of business records.

As far as comprehension processes are concerned, Edwards and Cantor (1991) have acknowledged the problematic use of jargon, and Sudman *et al.* (2000) have pointed to the problematic deviation of required economic concepts from accounting standards. The MIBSR model goes even further to explain that the errors may result from a broader issue of incomprehension of economic and accounting concepts or their confounding with other concepts.

The MIBSR model identifies several principles that help understand the underlying judgment processes in business surveys, which are consistent with examples manifesting the principles of continuity and consistency by Sudman, *et al.* (2000) and Willimack, Nichols and Sudman (2002), respectively. These principles may also reflect satisficing (Simon 1957) or inertia. The use of inappropriate principles, especially the principle of continuity, is particularly strengthened by the lack of survey feedback.

In the cognitive processes of responding, the MIBSR model exposes the problem of matching in business surveys, thus adding to the rounding error that Sudman *et al.* (2000) discuss. It also integrates different aspects of business sensitivity that Edwards and Cantor (1991) have discussed as part of the communication step, and Sudman *et al.* (2000) have discussed as part of the release step. The model treats them at the individual level where the editing occurs if the data are indeed sensitive.

4.4 Insights at the survey level

Previous models have concentrated on a single occurrence of the response process in a particular business survey, while the MIBSR model extends to several occurrences and several surveys. Among the many dimensions at the survey level, the study systematically analyzed the impact of recurrence and contact with the survey staff on the response process, which represents a further elaboration of specific instances already mentioned in previous models in the context of retrieval, such as rehearsal of the look-up (Edwards and Cantor 1991) or documentation of previous completions supporting retrieval (Sudman *et al.* 2000). In addition, the MIBSR model allows for the presence of a contagious effect transmitting the experience in one business survey to other business surveys.

5. Conclusion

Survey organizations usually have to set aside a considerable amount of resources for processing survey data because the processes of responding to survey questions in the businesses are not performed satisfactorily. The MIBSR model provides further evidence on how the processes are

carried out and what influences them. It offers insights into the business perspective, which are valuable for efficiently seeking solutions to improve the processes and, consequently, reduce or eliminate measurement errors. The model may also serve as a framework for the documentation and systematization of existing and future knowledge on the causes of measurement errors in business surveys. It may be used as a preceding step of empirical studies on measurement errors and for a consistent explanation of empirical findings. Future research should continue with the application of the qualitative research methods to the study of particular dimensions of the response process, other business participants besides respondents and other kinds of business surveys. It should also embark on quantitative modeling of the response process and verifying the effectiveness of suggested improvements with experiments. Last, it should look into the interactions with other kinds of nonsampling errors.

Acknowledgements

This article is an outcome of doctoral research. The author thanks the Statistical Office of the Republic of Slovenia for its co-operation and Lea Bregar (University of Ljubljana), Lars Lyberg (Statistics Sweden, Stockholm University) and Jaak Billiet (Catholic University of Leuven) for their guidance and support. I also thank the associate editor and anonymous referees for their helpful comments on an earlier version of this article.

References

Avison, D., and Elliot, S. (2006). Scoping the discipline of information systems. *Information Systems*: *The State of the Field*, (Eds., J.L. King and K. Lyytinen). Hoboken: John Wiley & Sons, Inc., 3-18.

Bavdaž, M. (2009). Conducting research on the response process in business surveys. *Statistical Journal of the IAOS*, 26, 1-14.

Beatty, P., and Herrmann, D. (2002). To answer or not to answer: Decisions processes related to survey item nonresponse. *Survey Nonresponse*, (Eds., R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little). New York: John Wiley & Sons, Inc., 71-85.

Biemer, P.P., and Fecso, R.S. (1995). Evaluating and controlling measurement error in business surveys. *Business Survey Methods*, (Eds., B.G. Cox *et al.*). New York: Wiley-Interscience, 257-281.

Edwards, W.S., and Cantor, D. (1991). Toward a response model in establishment surveys. *Measurement Errors in Surveys*, (Eds., P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz and S. Sudman). New York: Wiley-Interscience, 211-233.

Giesen, D., and Hak, T. (2005). The response process model in business surveys: Lessons learned by using a multi-method approach. *FCSM Conference Papers*, Federal Committee on Statistical Methodology.

Greenia, N., Lane, J. and Willimack, D. (2001). Perceptions of confidentiality protection at statistical agencies: Some evidence from data on businesses and households. *Statistical Journal of the United Nations ECE*, 18, 309-314.

Groves, R.M., Fowler, F.J., JR., Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R. (2004). *Survey Methodology*. Hoboken: Wiley-Interscience.

Hak, T., Willimack, D.K. and Anderson, A.E. (2003). Response process and burden in establishment surveys. *Proceedings of the Section on Government Statistics*, American Statistical Association, 1724-1730.

Hedlin, D., Dale, T., Haraldsen, G. and Jones, J. (2005). *Developing Methods for Assessing Perceived Response Burden.* Eurostat.

Jenkins, C.R., and Dillman, D.A. (1997). Towards a theory of self-administered questionnaire design. *Survey Measurement and Process Quality*, (Eds., L.E. Lyberg *et al.*). New York: Wiley-Interscience, 165-196.

Kvale, S. (1996). *InterViews*: *An Introduction to Qualitative Research Interviewing*. Thousand Oaks: Sage Publications.

Lorenc, B. (2006). Two topics in survey methodology: Modelling the response process in establishment surveys; inference from nonprobability samples using the double samples setup. Doctoral dissertation, Department of Statistics, Stockholm University.

Lorenc, B. (2007). Using the theory of socially distributed cognition to study the establishment survey response process. *Proceedings of the Third International Conference on Establishment Surveys*, *Montreal*, *Canada*, American Statistical Association, 881-891.

Morrison, R.L., Stettler, K. and Anderson, A.E. (2002). Using vignettes in cognitive research on establishment surveys. *International Conference on Questionnaire Development*, *Evaluation and Testing Methods*, *Charleston*, American Statistical Association.

Nichols, E.M., Murphy, E.D., Anderson, A.E., Willimack, D.K. and Sigman, R.S. (2005). Designing interactive edits for U.S. Electronic Economic Surveys and Censuses: Issues and guidelines. *Research Report Series* (*Survey Methodology* 2005-03), U.S. Census Bureau.

O'Brien, E.M. (2000). Respondent role as a factor in establishment survey response. *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, 1462-1467.

Phipps, P.A., Butani, S.J. and Chun, Y.I. (1995). Research on establishment-survey questionnaire design. *Journal of Business & Economic Statistics*, 13, 337-346.

Ponikowski, C.H., and Meily, S.A. (1989). Controlling response error in an establishment survey. *Proceedings of the Surveys Research Methods Section*, American Statistical Association, 258-263.

Ramirez, C. (1996). Respondent selection in mail surveys of establishments: Personalization and organizational roles. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 974-979.

Simon, H. (1957). *Models of man*: *Social and rational.* New York: John Wiley & Sons, Inc.

Snijkers, G., Onat, E. and Visschers, R. (2007). The annual structural business survey: Developing and testing an electronic form. *Proceedings of the Third International Conference on Establishment Surveys*, *Montreal*, *Canada*, American Statistical Association, 317-326.

Sudman, S., Willimack, D.K., Nichols, E. and Mesenbourg, T.L. (2000). Exploratory research at the U.S. Census Bureau on the survey response process in large companies. *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, 327-337.

Tomaskovic-Devey, D., Leiter, J. and Thompson, S. (1994). Organizational survey nonresponse. *Administrative Science Quarterly*, 39, 439-457.

Tourangeau, R. (1984). Cognitive science and survey methods. *Cognitive Aspects of Survey Methodology*: *Building a Bridge between Disciplines*, (Eds., T.B. Jabine, M.L. Straf, J.M. Tanur and R. Tourangeau). Washington, D.C.: National Academy Press, 73-100.

Tourangeau, R., Rips, L.J. and Rasinski, K.A. (2000). *The Psychology of Survey Response.* Cambridge, England: Cambridge University Press.

Willimack, D.K. (2003). Business respondents' perspectives on alternative employment arrangements and implications for employment statistics. *Proceedings of the Section on Government Statistics*, American Statistical Association, 4559-4570.

Willimack, D.K. (2007). Considering the establishment survey response process in the context of the administrative sciences. *Proceedings of the Third International Conference on Establishment Surveys*, *Montreal*, *Canada*, American Statistical Association, 892-903.

Willimack, D.K., Lyberg, L.E., Martin, J., Japec, L. and Whitridge, P. (2004). Evolution and adaptation of questionnaire development, evaluation, and testing methods for establishment surveys. *Methods for Testing and Evaluating Survey Questionnaires*, (Eds., S. Presser *et al*.). Hoboken: Wiley-Interscience, 385-407.

Willimack, D.K., and Nichols, E. (2001). Building an alternative response process model for business surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Willimack, D.K., Nichols, E. and Sudman, S. (2002). Understanding unit and item nonresponse in business surveys. *Survey Nonresponse*, (Eds., R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little). New York: John Wiley & Sons, Inc., 213-227.

Yin, R.K. (2003). *Case Study Research*: *Design and Methods.* Thousand Oaks: Sage Publications.

Survey Methodology, June 2010
Vol. 36, No. 1, pp. 95-109
Statistics Canada, Catalogue No. 12-001-X

Examining survey participation and response quality: The significance of topic salience and incentives

Lazarus Adua and Jeff S. Sharp [1]

Abstract

Nonresponse bias has been a long-standing issue in survey research (Brehm 1993; Dillman, Eltinge, Groves and Little 2002), with numerous studies seeking to identify factors that affect both item and unit response. To contribute to the broader goal of minimizing survey nonresponse, this study considers several factors that can impact survey nonresponse, using a 2007 Animal Welfare Survey Conducted in Ohio, USA. In particular, the paper examines the extent to which topic salience and incentives affect survey participation and item nonresponse, drawing on the leverage-saliency theory (Groves, Singer and Corning 2000). We find that participation in a survey is affected by its subject context (as this exerts either positive or negative leverage on sampled units) and prepaid incentives, which is consistent with the leverage-saliency theory. Our expectations are also confirmed by the finding that item nonresponse, our proxy for response quality, does vary by proximity to agriculture and the environment (residential location, knowledge about how food is grown, and views about the importance of animal welfare). However, the data suggests that item nonresponse does not vary according to whether or not a respondent received incentives.

Key Words: Survey nonresponse; Survey participation; Leverage-salience; Prepaid incentives; Item nonresponse; Missing data.

1. Introduction

Nonresponse bias has been a long-standing issue in survey research, as it affects all survey research regardless of mode (Nathan 2001). As a result, numerous studies have sought to identify factors that affect both item and unit response/nonresponse in various survey modes (Grove 2006; Trussell and Lavrakas 2004; Davern, Rockwood, Sherrod and Campbell 2003; Teitler, Reichman and Sprachman 2003; Singer, Van Hoewyk and Maher 2000; Singer, Van Hoewyk, Maher 1998; James and Bolstein 1992). While these studies have generated insightful and useful information about the factors that affect survey participation, questions about survey response still remain pertinent to the field of survey research in general and to our substantive work in particular. We are interested in expanding on the thoughts of Groves *et al.* (2000) by investigating whether specific characteristics of sampled units or demographic subpopulations in relation to a survey's topical context affect the response patterns. In our ongoing research assessing the general public's attitudes and behaviours related to the agricultural and environmental domain, we have become increasingly concerned about the level of survey participation and item nonresponse in distinct subpopulations. In our case, one concern is that unit and item nonresponse may vary among individuals or households that are more or less physically or socially proximate to the agricultural landscape, which is the focal area of our public opinion surveys.

To contribute to the broader goal of minimizing item and unit nonresponse and address some of our concerns, we reconsider several factors that can impact survey participation and item nonresponse. Specifically, we examine the effects of a survey's subject context (that is, its main focus) on survey participation and item nonresponse. We anticipate that participation in a survey will be systematically affected by how salient the survey's topic is to each sampled unit This expectation draws on the leverage-saliency theory (Groves *et al.* 2000), which anticipates that a variety of factors related to a survey's main features or features made prominent during survey administration might impact participation. Our research will also reconsider the effects of prepaid incentives on survey response. Given that offering incentives to sampled units has remained an enduring and widespread practice in the survey industry, we think it behoves survey researchers to periodically reassess the relationship between incentives and survey participation, using varying contexts. Such a continuous assessment of the utility of incorporating incentives into surveys is important because we cannot assume that incentives will always work as intended.

In the next section, we briefly describe the problem of survey nonresponse and then review research on how increasing the salience of some survey features and offering prepaid incentives affect participation and item non-response. The final two sections will cover the research design and results of the study.

1. Lazarus Adua, The Ohio State University, 330 Agricultural Admin Building, 2120 Fyffe Road, Columbus, OH43210, U.S.A. E-mail: adua.1@buckeyemail.osu.edu; Jeff S. Sharp, The Ohio State University, 254 Agricultural Admin. Building, 2120 Fyffe Road, Columbus, OH43210, U.S.A.

2. Survey nonresponse and potential consequences

Survey nonresponse describes the situation in which a sampled unit fails either to participate in the survey altogether (unit nonresponse) or to respond to one or more survey items (item nonresponse). Survey nonresponse has been a long-standing issue in survey research. Singer (2006) observes that "analysis of JSTOR statistical journals dates the first nonresponse article from 1945 and the *Public Opinion Quarterly* index's earliest reference is from 1948" (page 637). However, well-established and nascent survey projects alike are experiencing steadily declining response rates despite this awareness. For example, the University of Michigan's Survey of Consumer Attitudes (SCA) has witnessed a drop in response rate from about 72 percent in 1979 to about 60 percent in 1996 and a low of 48 percent in 2003 (Curtin, Presser and Singer 2005).

Survey nonresponse at both the unit and item levels obviously represents a major challenge to survey research, given its potential for generating nonsampling errors in parameter estimates (Brehm 1993; Dillman *et al.* 2002; Groves and Cooper 1998). For example, nonresponse may lead to biased point estimators, variance inflation for point estimators, and biases in estimators of precision (Dillman *et al.* 2002; Groves and Cooper 1998). Although unit and item nonresponse mean different things conceptually in the survey literature, their effects on a statistical estimate are generally the same (Groves, Fowler, Jr., Couper, Lepkowski, Singer and Tourangeau 2004).

While a number of recent studies suggest that low (unit) response rates may not have serious adverse effects on data quality (Curtin, Presser and Singer 2000; Keeter, Miller, Kohut, Groves and Presser 2000; Visser, Krosnick, Marquette and Curtin 1996), the fact still remains that unit nonresponse can have negative consequences for statistical estimates under certain circumstances. As a result, finding creative ways to increase response rates so that all types of sampled units are represented adequately in the sample remains a key goal in survey research. For item nonresponse, it may be true that advances in post-survey techniques for handling missing data, such as hot-deck and cold-deck imputations, mean imputation, multiple imputation, and multiple imputation and deletion, have made it possible to reduce the challenges this poses. However, the ideal situation and, in fact, a primary goal of survey design and implementation is to minimize item nonresponse to the greatest extent possible. This is because the norm in some fields, especially in microeconomics, is to use only the original data (Cameron and Trivedi 2009).

3. Making salient key features of a survey and survey participation

The extent to which a sampled unit views some features of a survey as more or less important affects the respondent's likelihood of participating in the survey (Groves *et al.* 2000). Groves *et al.* (2000) comment on the interviewing tactics of experienced interviewers, arguing that what interviewers actually do when they tailor their queries or remarks to the concerns of respondents is "to heighten the salience of some features of the request, those they judge will be favorably received by the household" (page 299). Building on Groves and Cooper (1998), Groves *et al.* (2000) propose what they call the *leverage-saliency theory* to explain how sampled units make the decision to participate or decline to participate in a survey. This theory essentially states that there are some attributes (leverage) of a survey that may be viewed negatively or positively by the respondent, and that how these attributes are made salient during the survey request process affects the likelihood of participation. If attributes viewed positively by a sampled unit (positive leverage) are made salient during the survey request, there is a higher chance that the respondent agrees to participate in the survey, all other things being equal. On the other hand, the likelihood of a sampled unit participating in a survey will be hurt if attributes that are viewed negatively by the respondent are made salient during the survey request.

Groves *et al.* (2000) empirically support this theoretical position. They present civic engagement (measured by community involvement) and incentives as leverages on survey participation, successfully showing that both attributes positively affect the likelihood of participation, with the effect of incentives diminishing among sampled units with higher civic engagement. In using civic engagement as a measure of a survey's leverage on sampled units, Groves *et al.* (2000) observe that leverage is not measured directly. Instead, it may be gleaned from some characteristic(s) of respondents in relation to the survey or its features, which may exert a positive or negative influence on the likelihood of participation. There is also evidence that when survey requests are tailored to the concerns of sampled units or to what they consider to be important, the likelihood of their participation is enhanced (Dillman 2000; Groves and Cooper 1998).

Based on the leverage-saliency theoretical proposition, we expect higher rates of participation from respondents whose characteristics make them more likely to view important attributes (leverage) of a survey positively. Correspondingly, we also expect those whose characteristics make them less likely to view such attributes positively to participate in the survey at lower rates. In our particular area

of research, we anticipate that sampled units' proximity to the agricultural and rural landscape (the contextual focus of our on-going survey) will affect participation in the survey and item nonresponse. This logic also applies to our expectations about respondents who claim greater knowledge of how food is produced and who also view animal welfare as important (a central sub-theme of this particular work). We thus draw from the leverage-salience theoretical proposition to propose the following hypotheses.

1. Our survey's focus on agriculture and the environment, which was made salient in its design, is expected to exert a positive leverage on respondents with greater social and physical proximity to agriculture and the rural environment (that is, those residing in more rural places). We thus hypothesize that participation rates will vary according to residential location.
2. We expect respondents with a closer proximity to agriculture and the rural landscape to be more diligent in completing the survey than those not in close proximity, as the former are more likely to be motivated by the survey's subject matter (that is, its positive leverage). We thus hypothesize that item nonresponse will vary by proximity to agriculture and the rural landscape.
3. Sampled units who have greater knowledge of how their food is grown as well as those who view animal welfare as important will have fewer item nonresponses. Presumably, such respondents will have a greater interest in the survey's focus on agriculture and the environment, and therefore exhibit more diligence in completing the survey.

4. Incentives and survey participation

The use of various forms of incentives, particularly prepaid (monetary) incentives, has become a common practice in survey research. While the practical rationale for offering incentives to sampled units is to encourage participation, the theoretical root of this practice is in part traceable to the *social exchange theory* (Dillman 1978). The social exchange theory assumes that people's actions are primarily motivated by the returns they expect or obtain from engaging in an activity (Weisberg 2005). Gouldner (1960) elaborates on the norm of reciprocity, which is related to the social exchange theory, observing that "insofar as men live under such a rule of reciprocity, when one party benefits another, an obligation is generated. The recipient is now *indebted* to the donor, and he remains so until he repays" (page 174). In Gouldner's view, the norm of reciprocity makes two demands on people: (1) people should help those who have helped them, and (2) people should not injure those who have helped them (Gouldner 1960, page 171).

Dillman (1978) uses the social exchange theory and particularly the social norm of reciprocity to argue that relatively small gestures (such as personalized letters, incentives, and reminder letters) can evoke reciprocation from sampled households in terms of inclination to participate in a survey. Also, Weisberg (2005) notes that social exchange is a theory that possibly explains the relationship between incentives and survey participation, observing that "[f]rom this perspective, giving the respondent a monetary incentive to participate in the survey can be seen as a kindness that evokes a norm of reciprocity" (page 165).

To devise 'ways and means' to bolster survey response rates as well as to test the social exchange theory in relation to incentive use in survey research, a number of experimental studies have examined the relationship between providing incentives to respondents and survey participation. While some of these studies have focused primarily on the effects of incentives on response rate and item nonresponse (Grove, Couper, Presser, Singer, Tourangeau, Acosta and Nelson 2006; Trussell and Lavrakas 2004; James and Bolstein 1992; Church 1993; Singer 2000; Yammarino, Skinner and Childers 1991; Fox, Crask and Kim 1988), others have examined the effects of incentives on respondent expectations and views about surveys (James and Bolstein 1990; and Singer *et al.* 1998). Consistent with the main proposition of the exchange theory and the norm of reciprocity, many of these studies report a positive relationship between incentives and response rates (Singer *et al.* 2000; Groves, Couper, Presser, Singer, Tourangeau, Acosta and Nelson 2006; Church 1993; Trussell and Lavrakas 2004; Goyder 1982; and Yu and Cooper 1983).

While many studies confirm the importance of incentives in encouraging survey participation, the empirically informed verdict on the relationship between incentives and survey participation is by no means unanimous. In a meta-analysis of experimental and quasi-experimental studies involving incentive conditions, Church (1993) reports that 1% of the studies utilized found no evidence of incentives affecting participation. Church also states that 10% of the 74 studies analyzed actually reported a negative relationship between the incentive conditions and survey participation. In fact, this reality partly prompted Groves *et al.* (2000) to propose the leverage-saliency theory to help explain why "incentives sometimes work" but "sometimes don't" (page 299). Given that findings related to the effects of incentives on survey participation are moderately mixed, as well as the fact that the subject matter of the survey we are studying differs from many previous studies, we find it necessary to

assess incentive effects on survey participation in conjunction with our examination of the relationship between agricultural proximity (our survey's contextual focus) and response. Also, we believe it is important to periodically assess the utility of using incentives in survey research, despite the fact that this subject has received a lot of attention in the past.

Another important incentive-related issue is the potential higher item nonresponse impacts of inducing reluctant respondents to participate in a survey (see Hansen 1980). The potential harm exists in that using persuasions such as incentives might elicit information from respondents who are careless or indifferent when answering questions, ultimately damaging the quality of the information obtained in this way (Singer *et al.* 2000). Owing to this concern, a number of studies have examined the relationship between incentives and item nonresponse, many of which suggest that incentives do not seriously harm response quality; that is, incentives do not generate higher item nonresponse (Singer *et al.* 2000; Singer *et al.* 1998; Shettle and Mooney 1999 and Davern *et al*. 2003). In fact, Singer *et al.* (2000) actually report that prepaid incentives help to reduce item nonresponse, an often-used measure of response or data quality. However, they also report that respondents who received incentives were more likely to give optimistic answers in some cases and be more pessimistic in others (involving different variables). In our case, a critical concern is that urban respondents induced to participate may provide lower quality data (as measured by nonresponse) than respondents more proximate to the agricultural and rural landscape.

In summarizing the review, we find that the research generally suggests that incentives help improve response rates in surveys, with little or no effect on item nonresponse. Although this is generally the case, some findings on the relationship do deviate from this expectation (Church 1993). Also, while many studies find that providing prepaid incentives does not affect item nonresponse, the work of Singer *et al.* (2000) suggests that providing incentives can compromise data quality via the mechanism of optimism or pessimism bias. Given these caveats, as well as the fact that most prior work on the relationship between incentives and survey participation was based on bivariate analysis (incentive and survey participation), we find it necessary to reconsider the impact of incentives on survey nonresponse while taking into account the effects of residential location in space and socioeconomic status. Thus, drawing from this literature on how incentives are related to survey participation and item nonresponse, we make the following hypotheses.

1. Respondents who received incentives will participate in the survey at higher rates than non-recipients, net the effects of proximity to the agricultural and rural landscape and socioeconomic status.
2. Incentives will be negatively related to item non-response. That is, surveys completed by respondents who received incentives will have fewer missing data points than those completed by respondents who did not receive incentives, controlling for the effects of respondents' proximity to the survey's subject and other covariates.

5. Study design

This paper is based on a survey of public views regarding food, agricultural and environmental issues, with a special focus on farm animal welfare. The target population of the survey was Ohio households. An initial sample of 3,000 respondents (along with their residential addresses) was drawn for the study via stratified random sampling: one-half (1,500) from Ohio's 22 core metropolitan counties and the second half (1,500) from the state's 66 metropolitan fringe or non metropolitan counties. The number of households in the core metropolitan counties differed from those in the metropolitan fringe or non metropolitan counties, making the sample a disproportionate random sample. To account for the unequal probability of selection across the two strata, we conducted weighted analysis for this paper.

The sample we used was obtained from Experian, a U.S.-based credit reporting bureau and private list vender. The sample was drawn from a sample frame (database) consisting of Ohio households along with their residential addresses. While we do not pretend that this sample frame covers all Ohio households, we believe that it is one of the most reliable and up-to-date lists and databases in the U.S. from which one can draw a sample. According to Experian, the database is updated monthly.

The survey followed a modified tailored design method (Dillman 2000) with up to four mailings sent to potential respondents during the spring of 2007. The first mailing was a pre-notification letter sent to each sampled unit, followed shortly by the survey packages. The third mailing was a reminder postcard sent to respondents thanking them for participating in the study or encouraging them to complete and return the survey if they had not yet done so. In the fourth mailing, replacement survey packages were mailed to respondents who had not returned completed questionnaires about 10 days after the postcard was mailed out. Of these four contacts with the respondents, three had information that focused specifically on the subject or topic of the survey. The pre-notification letter and the cover letters for the initial and replacement survey packages specifically conveyed to respondents the subject matter of the survey. Also, the graphics printed on the cover page of the survey

(images of farm animals) were selected to further convey this subject matter.

The addresses of sampled units were geo-coded and placed in a locational field (see details later in this section) to locate them geographically across the rural-urban continuum. This allowed us to conduct analyses of how sampled units' proximity to the agricultural landscape is related to their likelihood of participating in the survey. We recognize that some urban residents may have frequent social and physical interactions with agriculture and the rural landscape; however, this kind of interaction, along with its effects on support for agriculture and the environment, is highest among those residing in more rural and open country places (Freudenburg 1991; Sharp and Adua 2009). A randomized experiment involving incentives was also built into the survey. The first survey packages mailed to a randomly-selected half of the sampled units included $2.00 (two one dollar bills) incentives, while the other half of the sample received the same package but without any incentives. In doing this experiment, our pragmatic objective was to assess the effectiveness of our practice of enclosing modest cash incentives in survey packages to improve participation in our ongoing surveys of the Ohio public. Similar to Groves *et al.*'s (2000) expectations about the effect of community involvement on levels of participation, we also anticipated that households located in close proximity to agriculture and the rural landscape would participate at high levels in our study independent of the incentive, perhaps to the extent that a token financial incentive might be deemed unnecessary in future iterations of the survey.

5.1 Analytic strategy

Two sets of statistical analyses are conducted in this paper. The first set of analyses focuses on survey participation (response rate). First, we examine the proportion of successfully contacted sampled units who complete and return surveys by residential location along the rural-urban continuum, a proxy for geographic proximity to agriculture and rural areas of the state (an assumption we justify in a later section), and by incentive status. Following the American Association of Public Opinion Research's (AAPOR) 2008 guidelines for codes disposition, we defined successfully contacted sampled units as (i) those from whom we received completed surveys by the end of the data collection phase of the project, and (ii) those from whom we received neither a completed survey nor the survey package back from the United States Postal Service (USPS) as undeliverable. In our contract with the USPS, we requested that all mails that could not be delivered due to wrong address or absence of forwarding information be returned to us. The sampled units to which these undeliverable mails were addressed were classified as units we were unsuccessful in contacting. We also employ logistic regression to further analyze the likelihood of survey participation (coded 1 = responded; 0 = did not respond), using residential location along the rural-urban continuum and incentive status as the primary predictors, while simultaneously controlling for the effects of socioeconomic status at respondents' block group level as per the 2000 U.S. population census. We control for the effect of socioeconomic status because previous studies suggest it has some relationship with survey participation (Davern *et al.* 2003; Singer *et al.* 2000).

The second set of analyses focuses on item nonresponse. In this analysis, we conduct partial proportional ordered logistic regression analysis (generalized ordered logit) on the first two item nonresponse variables (0 = no missing items; 1 = some missing items; and 2 = numerous missing items), once again employing residential location along the rural-urban continuum and incentive status as the primary independent variables while controlling for the effects of several other variables. Generalized ordered logit (partial proportional odds) is employed rather than ordered logit because some predictors in these models violated the proportional odds assumption of ordered logistic regression. By using partial proportional odds modeling, we are able to constrain the relationship between those independent and dependent variables that met the proportional odds assumption of ordered logistic regression while allowing the relationships that failed this assumption to vary. To analyze the third item nonresponse variables, we employed logistic regression. This variable was recoded into a dichotomy (see the section on operationalization of variables for more details).

5.2 Operationalizing dependent variables

Survey Participation: Survey participation (response rate) is measured by computing the number of completed surveys received from respondents (eligible participating cases) as a proportion of the sampled units contacted successfully (all eligible cases). This measure of survey participation is in conformity with AAPOR guidelines for measuring response rates. Undeliverable surveys returned by the USPS without additional information, such as forwarding address or address correction, were treated as ineligibles. Cases for which we neither received completed surveys nor any other information about the cases from the USPS were treated as eligible based on the recommendation of the AAPOR's 2008 revised standard definitions of codes disposition and outcome rates. To conduct the logistic regression analysis of response likelihood, we coded all successfully contacted sampled units (eligible cases) as 1 (returned a completed questionnaire) or 0 (did not return a

completed questionnaire). We provide no descriptive statistics for this variable here as the analysis section, especially the marginals of the contingency tables, provides a good sense of the distribution of this variable.

Response quality: Response quality is measured by the occurrence of item nonresponse (see Davern *et al.* 2003; and Kaldenberg, Koenig and Becker 1994). To compute item nonresponse, missing data points for all respondents participating in the survey were summed across three subsets of items in the survey instrument to generate three item nonresponse variables: item nonresponse I, item nonresponse II and item nonresponse III. The item nonresponse I variable was created from items that, in our estimation, exerted comparatively the lowest cognitive demand on respondents, including such items as demographics and opinion questions that did not require very much introspection. The item nonresponse II variable was created from items that exerted comparatively higher cognitive demands on respondents than those used to create item nonresponse I, such as questions that required significant recall efforts and opinion questions that required a high level of introspection. The third variable is constructed from items that exerted comparatively the highest cognitive demand on respondents, such as knowledge questions and questions that required some understanding of concepts associated with animal husbandry.

In summing across these variables, we did not treat 'Don't Know' answers as item nonresponse, given that the survey had a couple of knowledge questions for which a 'Don't Know' response could be a legitimate answer. The item nonresponse variable also does not include "refused to answer" responses, as this option was not provided in questions used in the creation of the variables. We also excluded from these variables questions that respondents were directed to skip if they found them to be inapplicable.

Owing to the fact that the distribution of these variables was heavily skewed (see Table 1), the item nonresponse I and nonresponse II variables were regrouped into three ordinal categories (0 = no missing items; 1 = some missing items; and 2 = numerous missing items) and analyzed using generalized ordered logit. The first category (0) included cases without any item nonresponse, while the second category (1) included cases with between 1 and 9 incidences of nonresponse. The third category (2) included cases with 10 or more item nonresponses. For our analysis, we also regrouped the item nonresponse III variable into a dichotomy: 0 (no missing cases) and 1 (1 or more missing cases). This variable was regrouped differently from the first two because very few cases (only 19) satisfied the criteria for classification as "numerous missing cases" (Table 1). To verify whether our regrouping of these variables masked variances in item nonresponse within the groups (cases grouped together) that may be explained by our two independent variables (residential location, i.e. an indicator of interest in the survey topic, and incentives), we conducted a one-way analysis of variance for these grouped cases. Within these groups, none of the three item nonresponse variables varied significantly by residential location or incentives. Descriptive statistics for all three item nonresponse variables are reported in Table 1.

Table 1
Descriptive statistics for item nonresponse variables

	Item nonresponse I	Item nonresponse II	Item nonresponse III
Statistics before recoding			
N	971	971	971
Mean	3.11	2.34	1.6
Standard deviation	5.06	5.93	3.25
Minimum value	0	0	0
Maximum value	44	48	29
Statistics after recoding into groups			
Zero missing	30.07%	59.53%	54.69%
Some missing	62.31%	32.65%	43.36%
Numerous missing	7.62%	7.83%	1.96%

5.3 Operationalizing independent and control variables

Residential Location: The survey's focus on agricultural and environmental issues was made salient during the survey request (via the pre-notification letters, the cover letters and the design of the survey instrument), which can affect participation negatively or positively depending on each respondent's residential location along the rural-urban continuum. Residential location is an indicator of respondents' differentiated social and physical proximity to agriculture and the rural landscape. This is because proximity can increase the social and/or physical interactions with the subject. The association between proximity and environmental concern has been proposed and tested numerous times by social scientists (Dunlap and Heffernan 1975; Freudenburg 1991; Sharp and Adua 2009). We go a step beyond hypothesizing attitudinal differences associated with proximity and anticipate different levels of survey participation; indeed, we hypothesize that sampled units residing closer to agriculture and the rural landscape will participate in the survey at higher rates than those in core urban places. As a result, the subject matter of our survey is expected to serve as a positive leverage on sampled units residing closer to agriculture and the rural landscape. While this may not be a direct measure of leverage, it is consistent with Groves *et al.*'s (2000) suggestion that the leverage a given survey exerts on a sampled unit can be measured indirectly by relying on pertinent characteristics of the

sampled unit. In using the spatial residential characteristics of sampled units, we are relying on the fact that sampled units residing in more rural and open country areas have a higher likelihood of social and physical interaction with the agricultural and rural landscape than those in more urbanized places (see Table 2). In both 2006 and 2007, higher proportions of residents of exurban townships and rural areas (a combination of rural city/village and rural townships) visited a working farm than residents of core urban places, as shown in Table 2. We acknowledge that using information from our own respondents to show the association between residential location and visits to farms may be problematic. However, this information is corroborated by information from a different sample, the 2006 Ohio Survey.

To determine the residential location of the sampled units, each respondent's residential address was geocoded and assigned to one of four location fields – urban, suburban, exurban or rural – using ESRI's ArcView geocoding. Sampled units living in the exurban and rural fields were further distinguished as residing in either incorporated places (city/village) or township places (open country). This process of characterizing sampled units as living in urban, suburban, exurban, or rural places has previously been employed successfully in the field of regional science (Audirac 1999; Sharp and Clark 2008).

In this study, this variable has been grouped into five categories: (1) core urban, (2) suburban places, (3) exurban city/village, (4) exurban township and (5) rural places (cites/villages and townships). The ordering of the categories does not suggest a monotonic increasing order in terms of proximity to agriculture and the rural landscape between categories 1 and 5. Instead, this variable should be seen as a nominal variable with categories that can be grouped into blocks based on proximity to agriculture and the rural landscape: block 1(categories 1 and 2) has the lowest proximity, block 2 (category 3) has intermediate proximity and block 3 (categories 4 and 5) has the highest proximity. Between the blocks, the categories are monotonic increasing in terms of proximity to agriculture and the rural landscape, but within the blocks the pattern is less certain. Here, too, we provide no descriptive statistics for this variable as the analysis section provides an ample sense of how the variable is distributed.

Knowledge of Food Production and Support for Animal Welfare: Two other indicators of survey leverage used in the analysis are two survey items that measured sampled units' knowledge of how their food is produced and their views about the importance of animal welfare. The first asked, "How knowledgeable are you about how your food is grown? Please indicate on a scale of 1 to 7 your level of knowledge." This item had a mean of 4.47 and a standard deviation of 1.60. The second item asked, "Thinking about farm animals in general, how important is this issue to you? Please indicate on a scale of 1 (not important) to 7 (very important)." This item had a mean score of 4.50 and a standard deviation of 1.68. These two indicators are used in analyses pertaining only to the item nonresponse variables.

Incentive Status: Sampled units' incentive status (received versus did not receive incentive) is a primary independent variable in the regression models. Incentive status is dummy-coded as 0 (did not receive incentive) and 1 (received incentives). Again, we provide no descriptive statistics for this variable because the analysis provides a good sense of the variable's distribution.

Control Variables: Control variables operationalized in one or more of the analysis conducted in this study include *Age* (respondent's age as of his/her last birthday), *Education* (highest level of education completed), *Ethnicity* (white = 1; all others = 0) and *Gender* (male = 0 and female = 1), as well as the per capita and disposable median household income of each sampled unit's block group as per the 2000 population census. We control for the effects of these variables because previous studies suggest they can affect item nonresponse (Davern *et al.* 2003; Singer *et al.* 2000). Descriptive statistics for these purely control variables are shown in Table 3.

Table 2
Frequency of visiting or touring a working farm

	2006 Ohio Survey[a]			2007 Animal Welfare Survey[b]		
Residential location	**Never/ seldom**	**Occasional/ frequently**	**Total[c]**	**Never/ seldom**	**Occasionally/ frequently**	**Total[c]**
Core urban	90.4%	9.6%	100% (185)	81.0%	19.0%	100% (121)
Suburban place	87.5%	12.5%	100% (536)	83.7%	16.3%	100% (285)
Exurban city/village (Incorporated)	78.6%	21.4%	100% (217)	76.4%	23.6%	100% (124)
Exurban township (Unincorporated)	74.9%	25.1%	100% (434)	67.9%	32.1%	100% (264)
Rural place	73.1%	26.9%	100% (238)	70.6%	29.4%	100% (136)
Total	80.6%	19.4%	100% (1,610)	74.2%	25.8%	100% (930)

[a] Second-order corrected chi-square (3.61) = 43.3; $P = 0.0000$ (corrected for survey design effects)
[b] Second-order corrected chi-square (3.67) = 16.7; $P = 0.001$ (corrected for survey design effects)
[c] In parentheses are the total number of eligible cases from each residential category.

Table 3
Descriptive statistics for control variables

	Mean/percent	Standard deviation
Education:		
High school and lower	36.8%	-
Some college	32.3%	-
Bachelor's degree	13.7%	-
Grad/professional work & higher	17.2%	-
Gender:		
Male	48.2%	-
Female	51.8%	-
Ethnicity:		
White	91.7%	-
Non-white	8.3%	-
Age:	51.9	15.8
Block level mean household income, 2000	49,842.3	25,258.7
Block level median household income, 2000	42,616.3	16,728.6

6. Results

To evaluate survey participation, we use both bivariate analysis (contingency tables) and logistic regression modeling. For the contingency tables, we use Pearson chi-squared statistics corrected for survey design with Rao and Scott's (1984) second-order correction. We do this because survey design features such as stratification and clustering can affect tests of association (Lohr 1999). To limit the length of this paper, we follow a different analytical plan for the item nonresponse set of variables. For this set, we conduct only multivariate analysis (logistic regression). Moving straight to multivariate analysis allows us to examine the partial effects of the various predictors used in the models while keeping the paper brief.

6.1 Bivariate results for survey participation

The bivariate analysis suggests that survey participation varies significantly by proximity to the agricultural and rural landscape (residential location along the rural-urban continuum). As shown in Table 4, respondents residing in geographically more rural places (rural and exurban township residents) have higher rates of participating in the survey than those residing in geographically more urban places (core urban and suburban residents). The analysis also shows that those in the intermediate exurban incorporated places (cities and villages) were slightly more likely to participate than core urban residents. A second-order corrected chi-square test (Rao and Scott 1984) of the relationship between survey participation and residential location was significant ($\chi = 14.2$; df = 3.7; and $p = 0.003$).

Our analysis is consistent with previous studies, also finding that prepaid incentives significantly increase survey participation (Table 5). Despite the fact that the context of the survey used for our analysis differs markedly from previous studies examining the effects of incentives, we find that the response rate for successfully contacted incentive recipients was 43.7% compared with 26.9% for successfully contacted sampled units who did not receive prepaid incentives. The second-order corrected chi-square test of this bivariate relationship is also statistically significant ($\chi = 73.8$; df = 1; $p = 0.000$). In fact, our analysis suggests that eliminating incentives altogether substantially hurts participation rates for all categories of respondents regardless of proximity to the agricultural and rural landscape, although this effect is highest for residents in core urban places (Table 6). This finding provides support for our ongoing practice of using prepaid monetary incentives to help bolster our response rates with no discrimination between whether respondents reside in rural or urban locales. It also reaffirms the importance of incentives in survey research.

6.2 Logistic regression model for survey participation

Multivariate analysis further suggests that the likelihood of survey participation varies significantly by proximity to agriculture and the rural landscape, statistically holding constant the effects of incentive status (received versus did not receive incentive). Residents of suburban places, exurban townships, and rural places are significantly more likely to participate in the survey than residents of core urban places (Table 7). For example, residents of exurban townships and rural places have higher odds (0.60 log odds and 0.37 log odds, respectively) of participating than those of core urban places.

Table 4
Participation rate by residential location

Residential location	Responded	Did not respond	Total[a]
Core urban	29.5%	70.5%	100% (424)
Suburban place	32.6%	67.4%	100% (917)
Exurban city/village (Incorporated)	33.1%	66.9%	100% (379)
Exurban township (Unincorporated)	40.5%	59.5%	100% (684)
Rural place	35.8%	64.2%	100% (405)
Total	35.4%	65.6%	100% (2,809)

Second-order corrected chi-square (3.7) = 14.2; $P = 0.003$ (corrected for survey design effects)
[a] In parentheses are the total number of eligible cases from each residential category

Table 5
Survey response by incentive status

Incentive status	Responded	Did not respond	Total[a]
Incentive	43.7%	56.3%	100% (1,410)
No incentive	26.9%	73.1%	100% (1,401)
Tota1	35.4%	64.6%	100% (2,811)

Second-order corrected chi-square (1) = 73.8; $P = 0.000$ (corrected for survey design effects)
[a] In parentheses are the total number of eligible cases by incentive status

Table 6
Response rate by incentives and residential location along the rural-urban continuum

	Incentive recipients	Non-recipients of incentive	Response difference
Core urban	0.41	0.19	0.22
Suburban place	0.41	0.24	0.17
Exurban city/village (Incorporated)	0.39	0.27	0.12
Exurban township (Unincorporated)	0.48	0.31	0.17
Rural place	0.44	0.27	0.17
Total	0.43	0.26	0.17

Logistic regression analysis also seems to confirm our earlier finding that the likelihood of participating varies significantly by whether or not a sampled unit received incentives. Respondents who received incentives had higher odds (0.73 log odds) of participating in the survey than those who did not receive incentives, controlling for proximity to agriculture and the rural landscape as well as the gender (female = 1) of the householder randomly assigned as the preferred household member to complete and return the survey (Table 7).

Because socioeconomic status varies significantly by residential location across space (Lobao 1990) and affects survey response (Davern *et al.* 2003; Singer *et al.* 2000), we endeavored to control for the potential effects of per capita income and household income (socioeconomic status) on the likelihood of survey participation using hierarchical linear modeling (HLM). To do this, respondents were linked to their block groups and block group characteristics (specifically, block group per capita income and block group household median income) as per the 2000 U.S. population census. For the HLM analysis, we initially estimated a fully unconditional model (that is, an ANOVA) to determine whether the likelihood of survey participation varied significantly across the block groups. In hierarchical linear modeling, estimating a fully unconditional model (model without predictors at all levels of the analysis) is typically used to determine whether the dependent variable varies by the level two (or higher) unit of analysis, such as a neighborhood, block group or school district. This initial model (ANOVA) often helps researchers determine whether to proceed with multi-level analysis. Our initial HLM analysis (ANOVA) did not reveal any significant variation in the likelihood of survey participation across the block groups (tau = 0.04; $p = 0.493$). While this finding suggests the average probability of survey participation is about the same for all block groups despite their different per capita

and household disposable median incomes, we acknowledge potential instability in this HLM model given that sample cases per block group were generally low. This may have led to our finding of no significant variation in the likelihood of participation across the block group (potential Type II error). Despite this potential problem with our fully unconditional model, we did not proceed with the fully conditional multi-level analysis.

6.3 Logistic regression model for item nonresponse

As noted earlier in this section, our analysis of item nonresponse is limited to multivariate modeling, and we do this primarily to keep the paper brief while achieving our objective of assessing the partial effects of our main independent variables. The data suggest that the anticipated leverage of the survey's subject is only modestly related to item nonresponse. With respect to item nonresponse I (that is, the variables created from questions with the least cognitive demand on respondents in the survey), the analysis suggests that respondents in exurban township areas have lower item nonresponse (-0.74 log odds) than those residing in core urban areas, although this difference disappears at the higher values of this variable (Table 8, Columns 2 and 3). However, for item nonresponse II (the item nonresponse variables created from questions more cognitively demanding than those used in item nonresponse I) we find that residents of exurban townships and rural places are more likely to have higher item nonresponses (0.85 and 0.82 log odds, respectively) than residents of core urban areas (Table 8 Column 4). In terms of item nonresponse III (the item nonresponse variables created from the most cognitively demanding questions), the analysis did not reveal any significant difference by residential location, our proxy for level of interest in the survey's topic.

Supporting the anticipated effect of interest in a survey's topic on item nonresponse, the analysis also suggests that respondents' knowledge of how food is produced is significantly related to item nonresponse. In terms of item nonresponse II, the data shows that respondents who reported knowing how food is produced have lower log odds (-0.13) of item nonresponse than those who reported having less knowledge of how food is produced (Table 8, Column 4). This relationship is stronger at higher values of the variable: knowledge of how food is produced has lower log odds (-0.35) of item nonresponse when the category value shifts from 0 to 1 (Table 8 Column 5). This result suggests that the positive leverage of the survey's topic may have resulted in greater care in the completion of the survey among respondents with greater knowledge of how food is produced. We also find that respondents' views about the importance of animal welfare, a central subtheme of this particular survey, are positively related to item nonresponse (Table 8, Column 4). As shown in Table 8, a one unit increase in viewing animal welfare as important leads to a 0.09 unit increase in the log odds of item nonresponse (specifically item nonresponse II). This finding is inconsistent with our expectations.

In terms of the effects of incentives, we find no significant relationship between incentives and any of the three measures of item nonresponse (Table 8, Columns 2, 4 and 6), contrary to our expectation.

Table 7
Logistic regression[a] of likelihood of participation

	Log odds of participation	
	b	**Std. Error**
Incentive status		
Did not receive incentive (Ref)		-
Received incentive	0.73***	0.09
Residential location		
Core urban residents (Ref)	-	-
Suburban residents	0.27*	0.13
Exurban city/village residents	0.25	0.15
Exurban township residents	0.60***	0.13
Rural residents	0.37*	0.15
First option to respond (Female = 1)	-0.05	0.09
Model statistics		
Intercept	-1.42***	
Wald χ (df = 6)	93.25***	

Significance: ***< 0.001; **< 0.01; and *< 0.05

[a] In this model we tested for potential interaction effects between residential location and incentives. We found no evidence of such an interaction effect.

Table 8
Logistic regression models[a] for item nonresponse

	Item nonresponse I[b]		Item nonresponse II[b]		Item nonresponse III[c]
	No missing: log odds	Some missing: log odds	No missing: log odds	Some missing: log odds	Log odds
Incentive status					
Did not receive incentive	-				
Received incentive	0.16		0.10		-0.01
	(0.16)		(0.17)		(0.17)
Subject salience –Residential location					
Core urban residents	-				
Suburban residents	-0.14		0.54		-0.18
	(0.26)		(0.29)		(0.25)
Exurban city/village residents	-0.36		0.30		-0.24
	(0.31)		(0.34)		(0.29)
Exurban township residents	-0.74**	0.30	0.85**		-0.12
	(0.27)	(0.40)	(0.30)		(0.26)
Residents of rural places	-0.21		0.82**		0.08
	(0.3)		(0.31)		(0.29)
Subject salience –Food knowledge and animal welfare					
Knowledge about how food is produced	-0.07		-0.13*	-0.35***	-0.02
	(0.05)		(0.05)	(0.09)	(0.06)
Importance of animal welfare	0.10		0.09*		0.10
	(0.05)		(0.04)		(0.05)
Controls					
Education:					
High school and lower					
Some college	-0.79***		0.13		0.07
	(0.20)		(0.19)		(0.19)
Bachelor's degree	-1.08***		-0.32		-0.52
	(0.23)		(0.27)		(0.29)
Grad/professional work & higher	-0.99***		0.12		-0.38
	(0.24)		(0.24)		(0.24)
Age	0.03***		0.04***		0.03***
	(0.01)		(0.00)		(0.01)
Gender (Female = 1)	0.03		0.53**		0.21
	(0.17)		(0.17)		(0.17)
White	-0.38		-0.05		-0.51
	(0.32)		(0.28)		(0.32)
Model statistics					
Intercept	0.16	-4.36	-3.56	-5.07	-3.07
Wald chi-square[d]	85.80		93.25		54.87
N	828		828		828

Significance: ***< 0.001; **< 0.01; and *< 0.05
Standard errors shown in parentheses.
[a] We tested for potential interaction effects between residential location and incentives, between age and incentives and between ethnicity (white) and incentives in these models following Singer *et al.* (2000). We found no evidence of such interaction effects.
[b] The item nonresponse I and II models are partially constrained proportional odds logit models. This is because some of the predictors of these models violated the parallel lines assumption. These predictors were thus allowed to vary, while the remaining ones were constrained. William's (2006) gologit2 stata program code was used to estimate the model.
[c] This model is a logistic regression model with a binary dependent variable (variable recoded into two categories).
[d] Degrees of freedom are 14, 14, and 13 for the low cognitive, mid cognitive, and high cognitive models, respectively.

In terms of the control variables, we find that education is significantly related to item nonresponse, which is consistent with the earlier findings of Singer *et al.* (2000). In our case, respondents with some college work, a bachelor's degree, or some graduate/professional work have lower odds (-0.79, -1.08, and -0.99 log odds respectively) of missing cases for the survey questions with the lowest cognitive demand (item nonresponse I) than those with only a high school education or less (Table 8, Column 2). Surprisingly, item nonresponse related to the survey questions that were comparatively higher in cognitive demand (that is, item nonresponse II and item nonresponse III) did not differ by education (Table 8, Columns 4 and 6). We also find positive relationships between age and all three measures of item nonresponse (Table 8, Columns 2, 4, and 6), which is consistent with *Singer et al.* (2000). Equally consistent with the earlier work of Singer *et al.* (2000), the analysis reveals that female respondents are more likely to have missing data points than male respondents (Table 8, Column 4). However, the effect of gender on item nonresponse in our study is limited to those survey questions with a medium level of cognitive demand (the item nonresponse II variable).

7. Discussion and conclusions

In this study, we examined factors related to both unit and item nonresponse in survey research, focusing on interest in a survey's topic and prepaid incentives. The obvious reason for carrying out this analysis is the fact that nonresponse (unit or item) represents a major challenge to survey research given its potential for generating non-sampling errors in parameter estimates (Brehm 1993; Dillman *et al.* 2002; Groves and Cooper 1998). As previously noted, nonresponse can lead to biased point estimators, variance inflation for point estimators, and biases in estimators of precision (Dillman *et al.* 2002; Groves and Cooper 1998). Therefore, our primary goal is to provide information that will help researchers understand and deal appropriately with nonresponse, that is, minimize unit nonresponse and correctly understand and handle missing cases (item nonresponse).

Our analysis reveals that the likelihood of participation in this survey on agriculture and the environment varies significantly by sampled units' proximity to the agricultural and rural landscape (residential location). Our analysis is consistent with our first hypothesis and the theoretical proposition of leverage-salience, as we find that residents of exurban townships and rural places are all significantly more likely to participate in the survey than residents of core urban places. The pattern of relationships revealed in this analysis is most likely explained by the fact that respondents residing in exurban townships and rural places have a higher chance of interacting with the agricultural and rural landscape than those residing in core urban places (see Table 2). Thus, we suggest that respondents residing closer to the agricultural and rural landscape participated at higher rates in the survey due to the positive leverage of the survey's focus on the agricultural and environmental domain.

We also find some relationship between interest in the survey's topic (measured by proximity to the agricultural and rural landscape) and response quality (measured by item nonresponse). In support of our second hypothesis, modest evidence in this study suggests that item nonresponse varies by proximity to the agricultural and rural landscape. For item nonresponse I, the data suggest that residents of exurban township areas are less likely to have missing data points than residents of core urban places, whereas residents of both exurban townships and rural places are more likely to have missing data points for item nonresponse II. Missing cases associated with questions with the highest cognitive demand (item nonresponse III) did not vary by residential location (interest in the survey's topic). These findings suggest that residents of the more rural places (exurban townships and rural places) fare worse than those of core urban places when missing cases involve survey questions with a moderate level of cognitive demand. Although this result is intriguing, we are unable to explain why it is the case. One possible argument would be the educational difference between residents of core urban and rural places, but this study statistically controls for the effects of education. Further work certainly needs to be done on this subject.

Knowledge of how food is produced, another indicator of proximity to agriculture and the rural landscape, is negatively related to item nonresponse, which is consistent with our expectation (hypothesis 3) and the leverage-saliency theory. As the knowledge of how food is produced is related to the broader topic of the survey, we believe that making the survey's focus on agriculture and the environment salient in our request for participation in the survey may have generated higher diligence in questionnaire completion among respondents who knew or cared enough to know how food is produced. However, our analysis also suggests that support for animal welfare is positively related to item nonresponse, which is inconsistent with hypothesis 3. These findings highlight the need to look closely at factors related to a survey's topic as potential covariates of item nonresponse and its corollary, nonresponse error.

Although the survey used in this study focused on agriculture and the environment, our findings in relation to the survey's topic may have implications for surveys that focus on other sectors. There is reason to believe that unit

and item nonresponse can be affected by respondents' proximity to or level of interest in any survey topic or industry of focus, especially if this aspect of the survey is made salient during the request for participation. For example, if a survey focuses on the automotive industry and this feature is made salient during the request for participation, it is very likely that this information will affect the response pattern. In essence, these findings suggest that researchers designing surveys need to think critically about how the survey's subject context, such as the industry or sector on which it focuses, might affect participation from subpopulations within the sample list. While this generalization may be reasonable, we believe similar studies focusing on other sectors will be required before we can draw firm conclusions.

We next discuss the relationship between prepaid incentives on the one hand and survey participation and item nonresponse on the other. With respect to the relationship between incentives and response, our study suggests that prepaid incentives generally increase the likelihood of a respondent participating in a survey, even if proximity to agriculture and the rural landscape (the survey subject context) is taken into account. Our findings are consistent with hypothesis four and the previous literature (Singer *et al.* 2000; Groves 2006; Church 1993; Trussell and Lavrakas 2004; Goyder 1982; and Yu and Cooper 1983), as they show that recipients of prepaid incentives were significantly more likely to participate in the survey than non-recipients, controlling for other variables in the logistic regression model. The analysis demonstrates that eliminating incentives altogether hurts the likelihood of participation regardless of respondents' residential context. While we may not have overtly identified prepaid incentives with the leverage-saliency theory of Groves *et al.* (2000) in the earlier sections of our discussion for the sake of analytical convenience, our findings in relation to this variable also provide further empirical support for this theory. Our findings clearly suggest that token financial incentives enclosed with each survey package helped increase participation from both metropolitan and non-metropolitan areas of Ohio, although this effect was higher in the former. This result provides fresh justification for the widespread use of incentives to bolster response rates. As indicated earlier in this paper, the widespread use of prepaid incentives in surveys makes it necessary to periodically assess the utility of this practice. Our finding also suggests the need to check for potential response bias if incentives are provided to only a section of the sampled respondents, such as when prepaid incentives are targeted at those assessed as being less likely to participate.

In terms of the relationship between incentives and item nonresponse, we find no significant variation in missing data points between respondents who received monetary incentives and those who did not, contrary to our fifth hypothesis. This finding, which controls for the effects of residential location (proximity to the agricultural and rural landscape) and other pertinent variables, is consistent with the earlier work of Davern *et al.* (2003), who failed to find any relationship between incentives and the number of imputations for missing data points. Thus, while the use of monetary incentives correlates significantly with unit nonresponse (outright nonparticipation in a survey), we find no relationship between incentives and item nonresponse (failure to respond to some questions on a questionnaire). Thus, providing incentives to a respondent does not necessarily lead to greater diligence in survey completion.

The analysis revealed some interesting results with respect to the relationship between some of the control variables and item nonresponse. While education, age and gender were used in this study primarily as control variables, the fact that they were found to be significantly related to item nonresponse raises practical concerns about handling missing cases in survey data. Before choosing between the various techniques for handling missing cases (see Fuchs and Kenett 2007), analysts will need to check for potential nonresponse bias resulting from the effects of these variables, especially if they will be part of an analysis.

Acknowledgements

We thank Dr. Hebert Weisberg of the Political Science Department at The Ohio State University for commenting on an earlier draft of this paper. The survey upon which this paper is based was funded by the Ohio Agricultural Research and Development Center (OARDC), a research center within the College of Food, Agriculture, and Environmental Sciences at The Ohio State University. These acknowledgments notwithstanding, we are solely responsible for the content of this paper.

References

Audirac, I. (1999). Unsettled views about the fringe: Rural-urban or urban-rural frontiers. In *Contested Countryside*: *The Rural Urban Fringe in North America*, (Eds., O.J. Furuseth and M.B. Lapping). Brookfield, VT: Ashgate. 7-32.

Brehm, J. (1993). *The Phantom Respondent.* Ann Arbor: University of Michigan Press.

Cameron, A.C., and Trivedi, P.K. (2009). *Microeconometrics using Stata.* College Station, Texas: Stata Press.

Church, A.H. (1993). Estimating the effects of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly*, 57, 62-79.

Curtin, R., Presser, S. and Singer, E. (2000). The effects of response rate changes on the index of consumer sentiment. *Public Opinion Quarterly*, 64, 413-28.

Curtin, R., Presser, S. and Singer, E. (2005). Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Survey*, 69-1, 87-98.

Davern, M., Rockwood, T.H., Sherrod, R. and Campbell, S. (2003). Prepaid monetary incentives and item nonresponse in face-to-face interviews. *Public Opinion Quarterly*, 67, 139-147.

Dillman, D.A. (1978). *Mail and Telephone Surveys*: *The Total Design Method.* New York: John Wiley & Sons, Inc.

Dillman, D.A. (2000). *Mail and Internet Surveys*: *The Tailored Design Method*. New York: John Wiley & Sons, Inc.

Dillman, D.A., Eltinge, J., Groves, R. and Little, R. (2002). Survey nonresponse in design, data collection, and analysis. In *Survey nonresponse*, (Eds., Groves *et al.*). New York: John Wiley & Sons, Inc.

Dunlap, R.E., and Heffernan, R.B. (1975). Outdoor recreation and environmental concern: An empirical examination. *Rural Sociology*, 40, 18-30.

Fowler, Jr. F.J. (2002). *Survey Research Methods.* Third Edition. Thousand Oaks, CA: Sage.

Fox, R.J., Crask, M.R. and Kim, J. (1988). Mail survey response rate: A meta-analysis of selected techniques for inducing response. *The Public Opinion Quarterly*, 52-4, 467-491.

Freudenburg, W.R. (1991)**.** Rural-urban differences in environmental concern: A closer look. *Sociological Inquiry*, 61-2, 167-198.

Fuchs, C., and Kenett, R. (2007). Missing data and imputation. In *Encyclopedia of Statistics in Quality and Reliability*. (Eds., F. Ruggeri, R.S. Kenett and F. Faltin). Wiley. 1090-1099.

Gouldner, A. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review*, 25-2, 161-178.

Goyder, J.C. (1982). Further evidence of factors affecting response rates to mailed questionnaires. *American Sociological Review*, 47, 550-553.

Groves, R.M., and Cooper, M.P. (1998). *Nonresponse in Household Interview Surveys.* New York: John Wiley & Sons, Inc.

Groves, R.M., Singer, E. and Corning, A. (2000). Leverage-saliency theory of survey participation. *Public Opinion Quarterly*, 64, 299-308.

Groves, R.M., Fowler, Jr., F.J., Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R. (2004). *Survey Methodology.* Hoboken, NJ: John Wiley & Sons, Inc.

Groves, R.M., Couper, M.P., Presser, S., Singer, E., Tourangeau, R., Acosta, G.P. and Nelson, L. (2006). Experiments in producing nonresponse bias. *Public Opinion Quarterly*, 70-5, 139-147.

Groves, R.M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70-5, 646-675.

Hansen, R.A. (1980). A self-perception interpretation of the effects of monetary and nonmonetary incentives on mail survey response behavior. *Journal of Marketing Research*, 17, 77-83.

James, J.M., and Bolstein, R. (1992). Large monetary incentives and their effect on mail survey response rates. *Public Opinion Quarterly*, 56, 442-453.

Kaldenberg, D.O., Koenig, H.S. and Becker, B.W. (1994). Mail survey response patterns in a population of the elderly. *Public Opinion Quarterly*, 58-1, 68-76.

Keeter, S., Miller, C., Kohut, A., Groves, R.M. and Presser, S. (2000). Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly*, 64, 125-48.

Lobao, L.M. (1990). *Locality and Inequality*: *Farm and Industry Structure and Socioeconomic Conditions*. Albany, NY: The State University of New York Press.

Lohr, S.L. (1999). *Sampling*: *Design and Analysis*. Duxbury Press.

Nathan, G. (2001). Telesurvey methodologies for household surveys: A review and some thoughts for the future. *Survey Methodology*, 27-1, 7-31.

Rao, J.N.K., and Scott, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.

Sharp, J.S., and Adua, L. (2009). The social basis of agro-environmental concern: Physical versus social proximity. *Rural Sociology*.

Sharp, J.S., and Clark, J.K. (2008). Between the country and the concrete: Rediscovering the rural-urban fringe. *City & Community*. 7-1, 61-79.

Shettle, C., and Mooney, G. (1999). Monetary incentives in U.S. government surveys. *Journal of Official Statistics*, 15-2, 231-250.

Singer, E., Van Hoewyk, J. and Maher, M.P. (1998). Does the payment of incentives create expectation effects? *The Public Opinion Quarterly*, 62-2, 152-164.

Singer, E., Van Hoewyk, J. and Maher, M.P. (2000). Experiments with incentives in telephone surveys. *Public Opinion Quarterly*, 64, 171-188.

Singer, E. (2006). Introduction: Nonresponse bias in household surveys. *Public Opinion Quarterly*, 70-5, 637-645.

Teitler, J.O., Reichman, N.E. and Sprachman, S. (2003). Costs and benefits of improving response rate for a hard-to-reach population. *Public Opinion Quarterly*, 67, 126-138.

Trussell, N., and Lavrakas, P.J. (2004). The influence of incremental increases in token cash incentives on mail survey response. *Public Opinion Quarterly*, 68-3, 349-367.

Visser, P.S., Krosnick, J.S., Marquette, J. and Curtin, M. (1996). Mail surveys for election forecasting? An evaluation of the columbus dispatch poll. *Public Opinion Quarterly*, 60, 181-227.

Weisberg, H.F. (2005). *The Total Survey Error Approach*: *A Guide to the New Science of Survey Research.* Chicago and London: The University of Chicago Press.

Williams, R. (2006). Generalized ordered logit/Partial proportional odds models for ordinal dependent variables. *The Stata Journal*, 6-1, 58-82.

Yammarino, F.J., Skinner, S.J. and Childers, T.L. (1991). Understanding mail survey response behavior: A meta-analysis. *The Public Opinion Quarterly*, 55-4, 613-639.

Yu, J., and Cooper, H. (1983). A quantitative review of research design effects on response rates to questionnaires. *Journal of Marketing Research*, 20, 36-44.

Survey Methodology, June 2010
Vol. 36, No. 1, pp. 111-119
Statistics Canada, Catalogue No. 12-001-X

Evaluating within household selection rules under a multi-stage design

Tom Krenzke, Lin Li and Keith Rust [1]

Abstract

The 2003 National Assessment of Adult Literacy (NAAL) and the international Adult Literacy and Lifeskills (ALL) surveys each involved stratified multi-stage area sample designs. During the last stage, a household roster was constructed, the eligibility status of each individual was determined, and the selection procedure was invoked to randomly select one or two eligible persons within the household. The objective of this paper is to evaluate the within-household selection rules under a multi-stage design while improving the procedure in future literacy surveys. The analysis is based on the current US household size distribution and intracluster correlation coefficients using the adult literacy data. In our evaluation, several feasible household selection rules are studied, considering effects from clustering, differential sampling rates, cost per interview, and household burden. In doing so, an evaluation of within-household sampling under a two-stage design is extended to a four-stage design and some generalizations are made to multi-stage samples with different cost ratios.

Key Words: Intracluster correlation; Design effects; Multi-stage sampling.

1. Introduction

The 2003 National Assessment of Adult Literacy (NAAL), conducted for the National Center for Education Statistics, provided an indicator of the nation's progress in English literacy for researchers, practitioners, policymakers, and the general public. As in the 1992 National Adult Literacy Study (NALS), adults were assessed in households in prose, document and quantitative literacy. The booklet designs were based on the 1992 NALS to allow for the measurement of trends between 1992 and 2003.

In order to reduce the cost of interviewers traveling to households, the NAAL involved a stratified four-stage cluster design that resulted in 18,500 completed assessments administered to adults age 16 and older. In the NAAL, counties were grouped to form Primary Sampling Units (PSUs), which were stratified and selected in the first stage. In the second stage, Secondary Sampling Units (SSUs) were formed and selected within the sampled PSUs. The SSUs were individual census blocks, or groups of adjacent blocks with at least 60 households (HHs) formed within tract boundaries. Subsequently, households were selected within SSUs, and one sample person (1 SP) was randomly selected for household sizes up to $3 (B \leq 3)$, and two persons (2 SPs) were selected for household sizes greater than $3 (B > 3)$, where B denotes the number of eligible persons per household. This rule followed the within-household sampling approach used in the first cycle of NAAL (NCES 2001), conducted in 1992. An evaluation of the selection rule was conducted using the current US household size distribution and intraclass correlation coefficients computed from the 2003 survey. In doing so, an evaluation of within-household sampling under a two-stage design (Clark and Steel 2007) is extended to a four-stage design, as used in the NAAL survey and some generalizations are made to multi-stage samples with different cost ratios.

The data used for the evaluation include literacy measures from three scales derived from three types of literacy - prose, document, and quantitative. For more information about the NAAL types of literacy, refer to http://nces.ed.gov/NAAL/fr_tasks.asp. Two types of estimates are used; averages (*e.g.*, average prose literacy score) and percentage of adults at some level of literacy (*e.g.*, percentage *Below Basic* prose literacy). For a discussion of the literacy levels used in NAAL, see http://nces.ed.gov/NAAL/perf_levels.asp. In addition to the NAAL data, the evaluation also uses US sample data from the international Adult Literacy and Lifeskills (ALL), which was conducted by Statistics Canada. The US sample in 2003, sponsored by NCES, was part of a comparative study that measured the skills of adults in several countries. Similar to the NAAL, the ALL was a multi-stage clustered sample survey and measured prose and document literacy, as well as numeracy (OECD 2005). The NAAL sample was much larger (18,500 completes) than the ALL sample (3,400 completes), and the target population for NAAL included ages 16+ while the target population for ALL included 16 to 65 year olds. Table 1 provides a summary of each survey's design and structure.

1. Tom Krenzke, Statistical Group, Westat, Rockville, Maryland 20850. E-mail: tomkrenzke@westat.com; Lin Li, Statistical Group, Westat, Rockville, Maryland 20850. E-mail: linli@westat.com; Keith Rust, Statistical Group, Westat, Rockville, Maryland 20850. E-mail: keithrust@westat.com.

Table 1
Features of the NAAL and ALL surveys

Survey	Area sample	Completes	Data collection	Assessments	Ages	Within-HH sampling rule
NAAL	PSUs, SSUs households, Persons	18,500	Screener Interview Assessment	Prose Document Quantitative	16+	$B \leq 3$, $b = 1$ $B > 3$, $b = 2$
ALL	PSUs, SSUs, households, Persons	3,400	Screener Interview Assessment	Prose Document Numeracy	16-65	$B \leq 3$, $b = 1$ $B > 3$, $b = 2$

Note: PSU = Primary Sampling Unit, SSU = Secondary Sampling Unit, b = sample size, B = household size.

A discussion of the design considerations that helped form the evaluation of the within-household sampling rules is provided in Section 2. Section 3 discusses the computation of intra-household correlations under multi-stage sample designs and focuses on incorporating the clustering impact from the initial stages of sample selection when deciding on a within-household selection rule. An evaluation of selection rules was conducted using data from the in-person adult literacy surveys and the results are provided in Section 4. Finally, a brief summary is given in Section 5.

2. Design considerations

There are a number of factors that need to be considered when evaluating the within-households selection rules for surveys such as NAAL and ALL. The remainder of this section will discuss the impact of the following factors on within-household sampling: household burden, clustering persons within households, differential sampling rates, multi-stage sampling, cost considerations, computerized systems, domains of interest and household composition.

Household burden. For the adult literacy surveys, the interview and the assessment take about an hour and a half to administer in total. Therefore, one concern about selecting more than one person per household is the increase of burden to the household and the impact on response rates. However, there is no significant difference (0.05 significance level) in the refusal rates between 1- and 2-SP households in ALL and NAAL as shown in Table 2.

Clustering persons within households. Kish (1965) discusses the benefits of a cluster sample to a simple random sample. A cluster sample typically has a lower cost per person, however the unit variance is higher and it causes greater complexities in statistical analysis. Kish introduced the concept of a design effect (DEFF), which measures the increase in variance due to deviations from a simple random sample, such as clustering persons within households. Many surveys limit the selection to one sample person (SP) per household because of concerns over the increased clustering effect (*i.e.*, increasing effect on variance estimates) associated with multiple SPs per household. The DEFF due to clustering can be expressed as: $\text{DEFF}_{\text{clu}} = 1 + (\bar{b} - 1)$ Rho, where $\bar{b} = \sum(M_B / M) b_B$, M_B = number of households of size B, M = number of households, and b_B = sample size of persons within households of size B (Kish 1965). This DEFF component increases when the sample size within a household increases or when the value of the intracluster correlation (Rho) increases. As given in Cochran (1977), Rho can be approximated as:

$$\text{Rho} = 1 - \frac{\sigma_w^2}{\sigma^2},$$

where

$$\sigma_w^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} (y_{ij} - \bar{y}_{i.})^2 / (n - a),$$

and

$$\sigma^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} (y_{ij} - \bar{y}_{..})^2 / (n - 1),$$

where a is the number of sampled households, and b is the number of sampled persons per household. The DEFF due to clustering is examined further for different within-household sampling rules in the next section.

Differential sampling rates. A clustering effect is not the only factor that increases the variance. Increases in variance are also due to differential sampling rates (resulting in differential weights). Under a 1 SP per household strategy, the increase is directly related to the variation in household size since the sampling rate could vary from 1 out of 1 to 1 out of 7 or more. The DEFF due to differential sampling rates is expressed as: $\text{DEFF}_{\text{wgt}} = \sum(p_B / k_B) \sum(p_B k_B)$, where $p_B = N_B / N$, N_B = number of eligible persons in the population in households of size B, N = number of eligible persons in the population, and k_B = sampling rate within households of size B (Kish 1965). Under certain conditions, the overall DEFF can be expressed as the product of the clustering and differential sampling rate components: $\text{DEFF} = \text{DEFF}_{\text{clu}} \times \text{DEFF}_{\text{wgt}}$. Kalton, Brick and Lê (2005) suggest this product is applicable when the weights are random or approximately random.

Table 2
Refusal rates by 1- and 2-SP households for the adult literacy surveys

Survey	Subgroup	Refusal rate %
NAAL	1-SP households	16.3
	2-SP households	15.7
ALL	1-SP households	17.6
	2-SP households	16.2

Note: SP = sample person.

To arrive at a self-weighting sample, persons within households would need to be selected at a constant rate. However, a rate-based approach is not preferred in most surveys since it would result in walking away from a portion of single-person households and, thus, would increase the cost of the survey. We limit the alternative rules under consideration to those with a minimum of 1 SP per household. Out of concern for burdening households, the maximum sample size was set to two. The sampling rules under consideration are:

1. Take1: 1 SP no matter the household size.
2. Rule2: 1 SP for household sizes up to 2; otherwise 2 SPs are selected.
3. NAAL3: 1SP for household sizes up to 3; otherwise 2 SPs are selected.
4. Rule4: 1 SP for household sizes up to 4; otherwise 2 SPs are selected.
5. Frac5: take at least 1 SP, but no more than 2 SPs and the sample size is a fraction. That is, if the sample size for a household with two eligible persons is 1.6, then two persons are selected 60 percent of the time at random, and one person is selected 40 percent of the time.

While the Take1 approach does not attempt to reduce the DEFF due to differential sampling rates, it is not subject to a clustering impact. However, the other four approaches listed above provide a reduction in the differential sampling rate component while introducing a clustering effect. In the case of Frac5, under the assumption that π-weights are used, as assumed throughout this paper, the approach would result in the most reduction in the differential sampling rate component. The π-weights approach is based on the unconditional selection probability of the person within the household. If the actual sample size within a household is used in the form of ratio weights, the differential sampling rate increases the benefit is less clear and depends on Rho. Figure 1 illustrates the best options under a two-stage household design with fixed effective sample size of persons, without any cost considerations. The US national household size distribution from the 2007 Current Population Survey was used for this illustration. As shown in Figure 1, the fractional approach is the best rule for a wide range of values of Rho. The fractional approach can be programmed into a computerized system when enumerating and selecting household members (more discussion on computerized systems follows). If computerized systems are not available for screening, then the best approach for low values of Rho is the more clustered approach, Rule2; and the NAAL3 rule is best for Rho values greater than about 0.34.

Multi-stage sampling. For multi-stage area designs, the clustering impact of sampling within households is affected by the clustering due to PSUs and SSUs. As pointed out by Kish (1965), the clustering of households and persons within PSUs and SSUs increases the sampling variance (*i.e.*, units within PSUs and SSUs are more similar to each other). The incremental impact of clustering within households may be dampened by the domination of the PSU and SSU variance components (however, the magnitude of the impact will differ depending on the type of estimate and variable). That is, more persons within a household can be selected for surveys with a large amount of clustering due to the first two stages of sampling. Details of this distinction are provided in Section 3.

Cost considerations. The cost of screening a household in a 1 SP per household design versus the cost of interviewing/ assessing a second person in a household is investigated in an extensive analysis presented later .

Computerized systems. Computerized systems, such as Computer-Assisted Personal Interview (CAPI), have the capability of handling fractional sample sizes. That is, the random selection of 1 or 2 SPs given a pre-assigned fractional sample size can be programmed. Computerized systems also have the capability of sorting the list of eligible persons and selecting 2 SPs with a systematic random sample. Another benefit is that the selection program can be tested and validated prior to data collection.

Domains of interest. As mentioned earlier, optimal within household sampling depends on the magnitude of the clustering effect associated with the variable of interest. The clustering effect may be much smaller when the variable is associated with a subgroup of the population, rather than the entire population. For example, when a key reporting domain is gender in a survey of the adult population, the reporting category of males is likely to have an average of 1 SP per household and less likely to have 2 male SPs which would introduce a clustering effect. Therefore, when there are multiple domains of interest in a typical household, it is often beneficial to select more than 1 SP within a household. Refer to Mohadjer and Curtin (2008) for an example of design considerations for a survey with focus on multiple subgroups of the population.

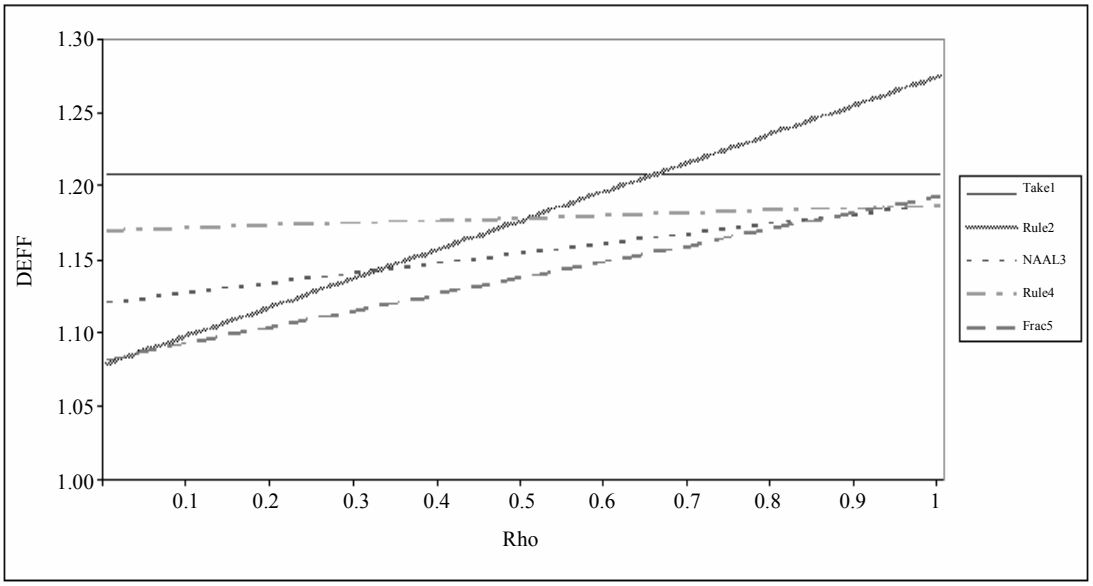

Figure 1 Initial analysis of within-household selection rules

Household composition. Lastly, one may want to consider the household composition and relationships of persons within a household when devising the selection rule. Table 3 displays values of Rho for various relationships between household members, for household with 2 SPs in the NAAL survey. Rho varies greatly by household member relationships. The relationships were derived from gender and age.

3. Estimation of *intra-household* Rho and DEFF under multi-stage sampling

The discussion about Rho thus far has been related to a two-stage design, but both NAAL and ALL have four stages of sampling. The total variance can be decomposed into four between-variance terms attributable to PSUs, SSUs, households and persons, as follows:

$$\sigma_T^2 = \sigma_{\text{PSU}}^2 + \sigma_{\text{SSU(PSU)}}^2 + \sigma_{\text{HH(SSU)}}^2 + \sigma_{\text{PERS(HH)}}^2.$$

As shown below, when applying a two-stage approach to estimate Rho for a four-stage sample design, the numerator not only contains the between household component, but also contains contributions from the between PSU and between SSU components inflating the values of Rho for our purpose.

$$\text{Rho} = 1 - \frac{\sigma_{\text{PERS(HH)}}^2}{\sigma_T^2} = \frac{\sigma_{\text{PSU}}^2 + \sigma_{\text{SSU(PSU)}}^2 + \sigma_{\text{HH(SSU)}}^2}{\sigma_T^2}.$$

Therefore, when evaluating rules for within-household sampling under a multi-stage design, we assume the PSU and SSU design will be the same in the future. This can be accomplished by limiting our focus to within SSU sampling. Therefore, the computation of Rho is contained within SSUs, that is, it is done in a compact manner without effect from the PSU and SSU components. We refer to this as the compact (*i.e.*, within SSU) Rho denoted by Rho^*, expressed as:

$$\text{Rho}^* = \frac{\sigma_{\text{HH(SSU)}}^2}{\sigma_{\text{HH(SSU)}}^2 + \sigma_{\text{PERS(HH)}}^2}.$$

Using the compact Rho^*, we now derive the estimated DEFF under a multi-stage sample design for the purpose of determining optimal within-household sample sizes. The variance of an estimate $(\hat{\theta})$ with b persons per household can be decomposed as:

$$\text{Var}(\hat{\theta}) = \frac{\sigma_{\text{PSU}}^2}{n_{\text{PSU}}} + \frac{\sigma_{\text{SSU(PSU)}}^2}{n_{\text{SSU}}} + \frac{\sigma_{\text{HH(SSU)}}^2}{n_{\text{HH}}} + \frac{\sigma_{\text{PERS(HH)}}^2}{bn_{\text{HH}}}$$

where, n_{PSU}, n_{SSU}, n_{HH} and bn_{HH} are the sample sizes of PSUs, SSUs, households and persons, respectively.

Table 3
Rho for NAAL assessment scores by household member relationships

Estimate	Siblings	Child-guardian	Married	Others
Number of households with 2 SPs	111	205	180	434
Average prose score	0.42	0.35	0.70	0.59
Average document score	0.40	0.27	0.72	0.54
Average quantitative score	0.46	0.36	0.63	0.56
Percentage Below Basic prose	0.52	0.41	0.79	0.67
Percentage Below Basic document	0.54	0.40	0.78	0.60
Percentage Below Basic quantitative	0.51	0.41	0.77	0.65

Then the DEFF due to clustering, relative to taking one person per household and bn_{HH} households is:

$$\mathrm{DEFF}_{clu}^{HH} = \frac{\frac{\sigma^2_{PSU}}{n_{PSU}} + \frac{\sigma^2_{SSU(PSU)}}{n_{SSU}} + \frac{\sigma^2_{HH(SSU)}}{n_{HH}} + \frac{\sigma^2_{PERS(HH)}}{bn_{HH}}}{\frac{\sigma^2_{PSU}}{n_{PSU}} + \frac{\sigma^2_{SSU(PSU)}}{n_{SSU}} + \frac{\sigma^2_{HH(SSU)}}{bn_{HH}} + \frac{\sigma^2_{PERS(HH)}}{bn_{HH}}}$$

$$= \frac{\frac{\sigma^2_{PSU}}{n_{PSU}} + \frac{\sigma^2_{SSU(PSU)}}{n_{SSU}} + \frac{1}{bn_{HH}}(\sigma^2_{HH(SSU)} + \sigma^2_{PERS(HH)} + (b-1)\sigma^2_{HH(SSU)})}{\frac{\sigma^2_{PSU}}{n_{PSU}} + \frac{\sigma^2_{SSU(PSU)}}{n_{SSU}} + \frac{1}{bn_{HH}}(\sigma^2_{HH(SSU)} + \sigma^2_{PERS(HH)})}$$

$$= \frac{\frac{bn_{HH}\left(\frac{\sigma^2_{PSU}}{n_{PSU}} + \frac{\sigma^2_{SSU(PSU)}}{n_{SSU}}\right)}{\sigma^2_{HH(SSU)} + \sigma^2_{PERS(HH)}} + (1 + (b-1)\,\mathrm{Rho}^*)}{\frac{bn_{HH}\left(\frac{\sigma^2_{PSU}}{n_{PSU}} + \frac{\sigma^2_{SSU(PSU)}}{n_{SSU}}\right)}{\sigma^2_{HH(SSU)} + \sigma^2_{PERS(HH)}} + 1}$$

$$= \frac{k^* + (1 + (b-1)\,\mathrm{Rho}^*)}{k^* + 1}$$

where,

$$k^* = \frac{bn_{HH}\left(\frac{\sigma^2_{PSU}}{n_{PSU}} + \frac{\sigma^2_{SSU(PSU)}}{n_{SSU}}\right)}{(\sigma^2_{HH(SSU)} + \sigma^2_{PERS(HH)})} = \frac{\frac{\sigma^2_{PSU}}{n_{PSU}} + \frac{\sigma^2_{SSU(PSU)}}{n_{SSU}}}{\frac{1}{bn_{HH}}(\sigma^2_{HH(SSU)} + \sigma^2_{PERS(HH)})}$$

Alternatively, DEFF_{clu}^{HH} can be expressed as:

$$\mathrm{DEFF}_{clu}^{HH} = 1 + \frac{(b-1)\,\mathrm{Rho}^*}{k^* + 1} = 1 + (b-1)\,\mathrm{Rho}^{**}$$

where,

$$\mathrm{Rho}^{**} = \frac{\mathrm{Rho}^*}{k^* + 1}$$

$$= \frac{\left(\frac{\sigma^2_{HH(SSU)}}{\sigma^2_{HH(SSU)} + \sigma^2_{PERS(HH)}}\right)\frac{1}{bn_{HH}}(\sigma^2_{HH(SSU)} + \sigma^2_{PERS(HH)})}{\frac{\sigma^2_{PSU}}{n_{PSU}} + \frac{\sigma^2_{SSU(PSU)}}{n_{SSU}} + \frac{\sigma^2_{HH(SSU)}}{bn_{HH}} + \frac{\sigma^2_{PERS(HH)}}{bn_{HH}}}$$

$$= \frac{\frac{\sigma^2_{HH(SSU)}}{bn_{HH}}}{\frac{\sigma^2_{PSU}}{n_{PSU}} + \frac{\sigma^2_{SSU(PSU)}}{n_{SSU}} + \frac{\sigma^2_{HH(SSU)}}{bn_{HH}} + \frac{\sigma^2_{PERS(HH)}}{bn_{HH}}}$$

$$= \frac{\sigma^2_{HH(SSU)}}{\frac{bn_{HH}\sigma^2_{PSU}}{n_{PSU}} + \frac{bn_{HH}\sigma^2_{SSU(PSU)}}{n_{SSU}} + \sigma^2_{HH(SSU)} + \sigma^2_{PERS(HH)}}$$

The Rho^{**} measure is a useful expression for the intra-household correlation under a multi-stage design, which is equal to Rho^* when $\sigma^2_{PSU} = \sigma^2_{SSU(PSU)} = 0$. The compact Rho^* measure is useful for evaluating optimal sample sizes while varying the variance ratio k^*. Note, however, that in general Rho^{**} is a function of n_{PSU}, n_{SSU} and the total sample size of persons, whereas Rho^* does not depend on these.

As shown in Table 4, the variance ratio k^*, which is the variance from the first two stages divided by the variance from the last two stages, for a one person per household design, ranges from 0.68 to 1.61 across types of assessments and estimates for the ALL survey.

Table 5 shows estimates for Rho (computed under a two-stage design assumption), the compact Rho^* and Rho^{**} (computed under a multi-stage design assumption where $k^* = 1$) for average NAAL and ALL literacy assessment scores. When including the clustering impact from the first two stages of the four-stage design, the values of the compact Rho^* and Rho^{**} are much smaller than Rho. For example, the two-stage Rho for the NAAL average prose score is 0.57 and the compact Rho^* is equal to 0.33 and Rho^{**} is equal to 0.17. The table also shows that values of the compact Rho^* for average scores are at about the same level for NAAL (range from 0.32 to 0.33) and ALL (range

from 0.29 to 0.39). There is some variation by the type of estimate as well; values of Rho^* for ALL are 0 to 0.2 lower for the percentage in Level 1 or 2 than for the average scores. Values of Rho^* can also vary by household size as shown in Figure 2 in Appendix A.

4. Evaluation and results

We compared the current sampling rules with optimal sampling rules by minimizing a variance-cost (VC) function, which is the product of the DEFFs (*i.e.*, variance increase) due to clustering and weighting, and a cost function that is used by Kish (1965):

$$\text{VC} = \text{DEFF}_{\text{clu}}^{\text{HH}*} \times \text{DEFF}_{\text{wgt}} \times n\left(c_p + \frac{c_{\text{HH}}}{\bar{b}}\right),$$

where c_p = cost per added person and c_{HH} = cost per added household. Note that $n/\bar{b}$ represents the number of sampled households. To account for the differential clustering effects for each household size B, we replace $\text{DEFF}_{\text{clu}}^{\text{HH}}$ with:

$$\text{DEFF}_{\text{clu}}^{\text{HH}*} = \frac{k^* + \sum_B \frac{M_B}{M}(1 + (b_B - 1)\text{Rho}_B^*)}{k^* + 1}$$

where Rho_B^* is computed as described in Appendix A.

Note that the VC function represents the additional cost of increasing the overall sample size to offset the increase in variance due to the DEFF components. Table 6 provides the results for optimal integer solutions as computed by a computational algorithm which is described in Appendix B. The table shows that as the cost ratio increases from 0.5 to 1 for $k^* = 1$, we would want to take more persons per household, that is, 2 out of 2 instead of 1 out of 2. As the variance ratio goes from 1 to 3 for optimal integer solutions, the only change is for household size of 2 and cost ratio of 0.5. That is, when the variance ratio is equal to 3, it is beneficial to take 2 out of 2 instead of 1 out of 2.

Table 6 also gives the results when fractional sample sizes are allowed. The variance and cost ratios for NAAL and ALL tend to be about 1, where it appears that selecting 1 out of 1, 1.6 out of 2, and 2 otherwise is the best rule. The effects of cost and variance ratios are clearer under the fractional sample sizes when compared to the integer solutions.

If the cost of conducting a screener is small in relation to the cost of interviewing, then variances can be reduced using the fractional walk-away approach. Table 6 shows optimal walk-away sample sizes. Under this approach, for example, a sample size of 0.9 indicates that we walk away from 10 percent of the households where $B = 1$. If the cost of screening is a very small portion of the cost of interviewing, then the optimal design may involve walking away from many more households.

Under the likely NAAL/ALL parameters for cost ratios ($C_{\text{HH}}/C_p = 1$) and variance ratios ($k^* = 1$), when compared to the Take1 approach, the VC function can be reduced by about 9 percent by using the NAAL/ALL sampling rule, 19 percent by using the optimal integer solution, 20.4 percent using the optimal fractional solution, and 20.6 using the optimal walk-away approach. In general, the gains from deviating from the Take1 approach grow as the cost per additional households (*i.e.*, screening) increases. The average cluster sizes for each approach are given in Table 7. For the NAAL and optimal integer rule, the average cluster size indicates the percentage of households with 2 SPs. For example about 6 percent of the households would have 2 SPs under the NAAL3 strategy.

Table 4
Values of k^* for the ALL sample

ALL estimate	k^*
Average prose score	0.95
Average document score	1.56
Average quantitative/numeracy score	1.13
Percentage in Level 1 or 2 prose	0.68
Percentage in Level 1 or 2 document	1.61
Percentage in Level 1 or 2 numeracy	1.10

Table 5
Values for Rho, Rho^*, and Rho^{**} for literacy assessment scores

	Rho		Rho^*		Rho^{**}	
Estimate	NAAL	ALL	NAAL	ALL	NAAL	ALL
Number of households with 2 SPs	930	162	930	162	930	162
Average prose score	0.57	0.60	0.33	0.38	0.17	0.19
Average document score	0.53	0.50	0.33	0.29	0.17	0.15
Average quantitative/numeracy score	0.54	0.58	0.32	0.39	0.16	0.20
Percentage Below Basic(NAAL)/Level 1or 2 (ALL) prose	0.65	0.44	0.42	0.28	0.21	0.14
Percentage Below Basic (NAAL)/Level 1 or 2 (ALL) document	0.61	0.37	0.39	0.28	0.20	0.14
Percentage Below Basic quantitative (NAAL)/Level 1 or 2 (ALL) numeracy	0.62	0.36	0.40	0.17	0.20	0.09

Note: Rho^{**} is computed assuming $k^* = 1$.

Table 6
Optimal expected number of persons per *household* by type of person sampling method and *household* size (B)

		Person Sampling Method											
		Integer				Fractional				Walk-away			
k^*	C_{HH} / C_p	$B = 1$	$B = 2$	$B = 3$	$B = 4$	$B = 1$	$B = 2$	$B = 3$	$B = 4$	$B = 1$	$B = 2$	$B = 3$	$B = 4$
1	0.5	1	1	2	2	1	1.4	2	2	0.6	1.3	2	2
1	1	1	2	2	2	1	1.6	2	2	0.9	1.6	2	2
1	2	1	2	2	2	1	1.9	2	2	1	1.9	2	2
3	0.5	1	2	2	2	1	1.6	2	2	0.8	1.5	2	2
3	1	1	2	2	2	1	1.8	2	2	1	1.8	2	2
3	2	1	2	2	2	1	2	2	2	1	2	2	2

Table 7
Percent reduction of NAAL3 and optimal solutions from Take1 strategy and average cluster sizes

		Percentage reduction from Take1 strategy				Average cluster sizes			
k^*	C_{HH} / C_p	NAAL3	Integer	Fractional	Walk- away	NAAL3	Integer	Fractional	Walk- away
1	0.5	8.2	13.0	15.8	18.0	1.06	1.18	1.38	1.21
1	1	9.1	19.2	20.4	20.6	1.06	1.68	1.48	1.45
1	2	9.9	26.1	26.1	26.1	1.06	1.68	1.63	1.63
3	0.5	8.6	17.3	18.7	19.0	1.06	1.68	1.48	1.37
3	1	9.5	23.7	23.9	23.9	1.06	1.68	1.58	1.58
3	2	10.4	30.2	30.2	30.2	1.06	1.68	1.68	1.68

Lastly, a sensitivity analysis was conducted by varying the values of Rho*. A regression model was fit on the percentage reduction from the Take1 strategy of the VC function, with the independent variables being the approach (NAAL3, integer, fractional, walk-away), cost ratio (0.1, 0.5, 1, 2, 10), variance ratio (1, 3, 5) and Rho* (+/- 0.1). For the range of data, Rho* had a limited impact (parameter estimate -7.4 with an associated standard error of 4.5) on the percentage reduction of the VC function, while the other factors had more of an impact.

5. Summary

Several design considerations were taken into account when evaluating the within - household selection rule for the NAAL and ALL surveys, including taking into account clustering effects from initial stages of sampling. To facilitate the evaluation, we formulate a way to incorporate PSU and SSU variance contributions into the computation of the DEFF due to clustering and the intra-household correlation when deciding how many persons and how many households to select in a multi-stage sample design. In doing so, we introduce compact Rho* measure, which is computed within the SSU so it is not impacted by the PSU and SSU variance components. This is useful when determining the DEFF due to clustering within households, while varying the contribution to the total variance from the PSU and SSU stages of selection in multi-stage sample designs. The measure Rho** is introduced as an expression for the intra-household correlation under a multi-stage design, taking into consideration the contribution to total variance from the first two stages of selection.

In addition, a computational algorithm was developed to compute optimal sample size solutions, incorporating the DEFFs due to clustering, differential sampling rates, and costs.

In general, the main factors on the percentage reduction of the VC function from the Take1 approach are the level of dominance from the PSU and SSU variance components in multi-stage sampling, the cost ratio and the rule used. For the range of data evaluated, Rho* had limited impact on the reduction in VC from the Take1 approach. In general, the NAAL rule improves on the widely-used Take1 approach. The optimal integer rule improves on the NAAL rule. However, the optimal fractional rule has limited gains over the optimal integer rule. The optimal walk-away rule has gains over the other rules for lower cost ratios. Lastly, when the first two variance components dominate and cost ratio is high, then the integer, fractional and walk-away rules are essentially the same.

Acknowledgements

The authors acknowledge valuable contributions by Leyla Mohadjer and Bob Fay.

Appendix A

Estimates of Rho* by household size

Survey estimates are not attainable for Rho* by household size since only 1 SP was selected for household size of 3 or less and since the sample size was too small to create estimates for each household size of 4 or more. Therefore, estimates of Rho* by household size are modeled using Census data. Figure 2 shows Rho* on the y-axis and household size on the x-axis. The upper line is from the US Census public-use microdata sample (PUMS) file for education attainment for ages 25+. The upper line shows that education attainment is more similar among households with two adults, perhaps more likely to be married couples. It shows a drop off when going from two to three adults. We captured the variation in households size by computing the ratio of Rho* for the NAAL prose literacy scores to the Rho for the Census PUMS education attainment among households with $B>3$ and applying the ratio to the PUMS Rho across all household sizes. The resulting values are the estimates of compact Rho^*_B for $B=1,\ 2,\ \ldots 11$.

Appendix B

Computational algorithm

A computational algorithm was developed to arrive at optimal within-household sample sizes for each household size B. The algorithm was constructed to generate optimal integer or fractional solutions that capture the effects of clustering, differential sampling rates and cost, under the constraints of at least one selected person per household and no more than 2. Here are the steps of the algorithm (all processing runs converged within four iterations):

- Initialize by setting $b=1$ for all values of B (Take1).
- Compute $\text{DEFF}_{\text{clu}}^{\text{HH*}}$, DEFF_{wgt}, c_p, c_{HH}, and VC(0).
- Do $I=1$ to 5.
  - Do $B=1$ to 11.
    - Compute $\text{DEFF}_{\text{clu}}^{\text{HH*}}$, DEFF_{wgt}, c_p, c_{HH}, and VC for all $1\le b_B\le 2$, given the set of b_B, for all $B'\ne B$.
    - Identify the b_B with the smallest value of VC.
  - End.
  - If $\text{VC}(I)=\text{VC}(I-1)$ then stop.
- End.

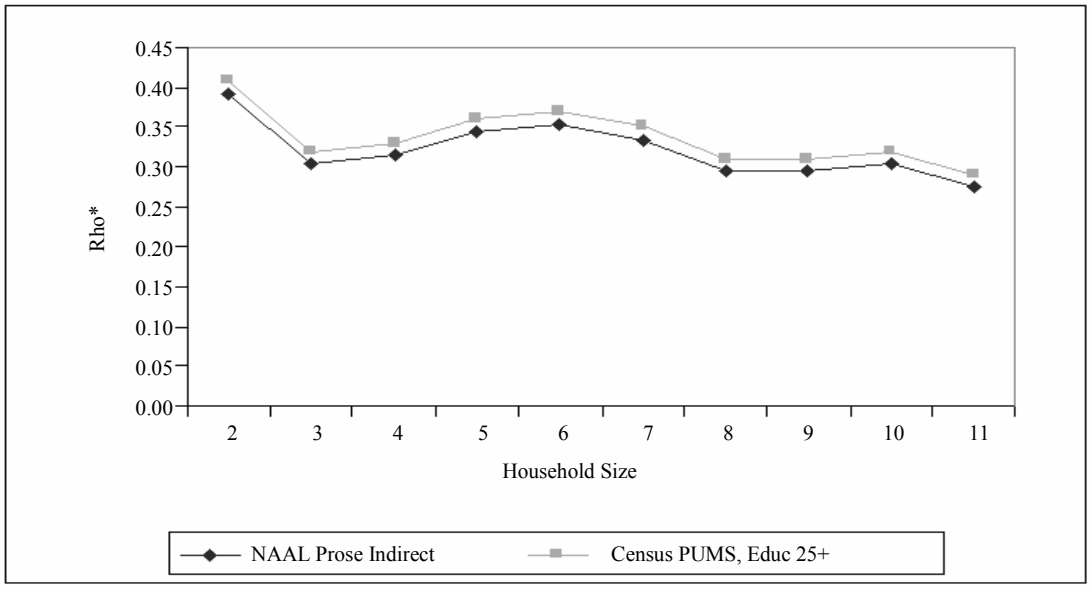

Figure 2 Estimates of Rho* for NAAL by household size

References

Clark, R.G., and Steel, D.G. (2007). Sampling within Households in Household Surveys. *Journal of the Royal Statistical Society*, Series A, 170, 63-82.

Cochran, W.G. (1977). *Sampling Techniques*. 3rd Ed. New York: John Wiley & Sons, Inc.

Kalton, G., Brick, J.M. and Lê, T. (2005). Estimating Components of Design Effects for Use in Sample Design, Household Sample Surveys in Developing and Transition Countries, Chapter VI United Nations, New York, 95-121.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Mohadjer, L., and Curtin, L.R. (2008). Balancing sample design goals for the National Health and Nutrition Examination Survey. *Survey Methodology*, 34, 1, 119-126.

NCES (2001). Technical Report and Data File User's Manual For the 1992 National Adult Literacy Survey. U.S. Department of Education, National Center for Education Statistics, Washington, DC.

OECD (2005). Learning a Living: First Results of the Adult Literacy and Life Skills Survey. Organisation for Economic Co-operation and Development, Paris. Statistics Canada, Ottawa.

2010 International Methodology Symposium
Statistics Canada
October 26-29, 2010
Ottawa, ON, Canada

Social Statistics: The Interplay among Censuses, Surveys and Administrative Data

Statistics Canada's 2010 International Methodology Symposium will take place at the Crowne Plaza Hotel, located in the heart of downtown Ottawa, from October 26-29, 2010.

The Symposium will be titled **"Social Statistics: The Interplay among Censuses, Surveys and Administrative Data"**. Members of the statistical community, such as those from private organizations, governments, or universities, are invited to attend, particularly if they have a special interest in statistical or methodological issues resulting from the use of multiple sources of data (censuses, sample surveys or administrative data).

The first day will consist of workshops, while the following days will consist of both plenary and parallel sessions covering a variety of topics. Additional research and results may be presented via poster sessions.

The presentations will be related to the methodological aspects of using multiple sources of data. Topics may include:

- Sampling Frames and Sample Design
- Coordinating Samples
- Content and Questionnaire Design
- Data Collection Methods and Acquisition of Administrative Data
- Supplementing Survey Data with Administrative Data
- Administrative Data for Direct Estimation
- Statistical Databases from Administrative Data (e.g., Population Registers)
- Imputation
- Weighting and Estimation
- Dissemination and Data Access
- Record Linkage Techniques
- Record Linkage Software
- Measurement Errors
- Response Burden
- Treatment of Nonresponse
- Confidentiality, Privacy and Ethical Issues
- Small Area Estimation

Visit our Internet site regularly to obtain further details about the program, workshops, registration, accommodation, tourism information and more at

http://www.statcan.gc.ca/conferences/symposium2010/index-eng.htm

Statistics Canada Statistique Canada

Canada

FOURTH INTERNATIONAL CONFERENCE ON ESTABLISHMENT SURVEYS (ICES IV) PLANNED FOR 2012

Planning is underway for the Fourth International Conference on Establishment Surveys (ICES IV). If you've attended any of the past conferences, you know how invaluable they have been to the literature and practice of establishment surveys. If you are newer to the establishment survey field, you will find the conference especially rewarding. Since the last ICES held in 2007, many new techniques have been developed by practitioners around the world. A major strength of the conferences is the strong international presence, both in the program development and attendance. Over 400 people from 94 countries attended ICES III. On June 11-14 2012, survey practitioners from government agencies, academia, private sector and more will gather at the Sheraton Centre Montreal in Quebec, Canada for ICES IV and continue the tradition of sharing innovative techniques and best practices to address common issues.

Sponsorship of the meetings is being provided by the American Statistical Association, ASA Section on Survey Research Methods, ASA Section on Government Statistics, International Association of Survey Statisticians, and the Statistical Society of Canada. Administrative support for ICES IV will be provided by the American Statistical Association, similar to previous ICES meetings. Also, many other organizations and government agencies are or will be providing support for the conference.

With the support of these many great organizations and the diverse gathering of individuals involved in establishment surveys, we anticipate that ICES IV will prove to be another fruitful conference in the valuable ICES series. So, save the date, **June 11-14, 2012**, and join practitioners from around the globe in **Montreal, Canada**! You can participate in the growing ICES IV program discussing current issues, future vision, and cutting-edge methods in surveying businesses, farms and institutions. Expect updates on participation and program details to ICES IV through this newsletter and the upcoming ICES IV website. Inquiries may be directed to ices4@amstat.org.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 25, No. 4, 2009

All inquires about submissions and subscriptions should be directed to jos@scb.se

CONTENTS TABLE DES MATIÈRES

Volume 37, No. 4, December/décembre 2009

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 38, No. 1, March/mars 2010

GUIDELINES FOR MANUSCRIPTS

Before finalizing your text for submission, please examine a recent issue of *Survey Methodology* (Vol. 32, No. 2 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A pdf or paper copy may be required for formulas and figures.

1. Layout

1.1 Documents should be typed entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The documents should be divided into numbered sections with suitable verbal titles.
1.3 The name (fully spelled out) and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", *etc*.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (*e.g.*, w, ω; o, O, 0; l, 1).
3.6 Italics are used for emphasis.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles that are as self explanatory as possible, at the bottom for figures and at the top for tables.

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, *e.g.*, Cochran (1977, page 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

6. Short Notes

6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.