Performance of Generalized Regression Estimator and Raking Estimator in the Presence of Nonresponse

Daifeng Han[1], Richard L. Valliant[2], Jill M. Montaquila[3], Keith Rust[4]

[1]Westat, 1600 Research Blvd, Rockville, MD 20850
[2]Joint Program in Survey Methodology, University of Maryland, 1218 LeFrak Hall, College Park, MD 20742
[3] Westat, 1600 Research Blvd, Rockville, MD 20850
[4]Westat, 1600 Research Blvd, Rockville, MD 20850

Abstract

Calibration weighting is widely used to decrease variance, reduce nonresponse bias, and improve the face validity of survey estimates. In the purely sampling context, Deville & Särndal (1992) demonstrate that many alternative forms of calibration weighting are asymptotically equivalent, so the generalized regression (GREG) estimator can be used to approximate some general calibration estimators with no closed-form solutions. It is unclear whether this conclusion holds when nonresponse exists and single-step calibration weighting is used to reduce nonresponse bias (i.e., calibration is applied to the basic sampling weights directly without a separate nonresponse adjustment step).

In practice, poststratification (as a special form of the GREG estimator) and raking (as an example of general calibration estimators) are commonly used calibration approaches, but decisions between these estimators are often made ad-hoc based on sample sizes and availability of external data. In this paper, we compare the performance of these estimators by examining their biases, variances, and effective coverage of the confidence intervals. The theoretical work and simulation study demonstrate the need to consider models for both the outcome variable and the response pattern. The model supporting the typical application of raking has main effects only while poststratification (and more general forms of GREG) can include interactions. A framework involving both design-based and model-based thinking is developed to simultaneously evaluate the impact of sampling, outcome variable structure, and nonresponse mechanism.

Since survey practitioners often lack the knowledge of the outcome variables and nonresponse mechanism in real-world surveys, we also develop a diagnostic that helps gauge the potential consequence of choosing an inappropriate calibration estimator. The results of this research will provide guidelines for choosing between the commonly used calibration estimators.

Keywords: calibration, GREG, poststratification, nonresponse adjustment

1. Introduction

Calibration weighting has originally been developed as a method for reducing sampling errors while retaining randomization consistency. Deville and Särndal (1992) introduce calibration estimators using the distance function approach. Later work by Särndal (2007) points out that there are two different approaches to take account of auxiliary information in the estimation – "calibration approach" and "regression approach". The two approaches generate the same estimator, the generalized-regression (GREG) estimator, in the situation where the general linear squares (GLS) distance function is used in the calibration approach and linear regression model is used in the regression approach. For the purpose of comparison, we use the term "general calibration estimators" to refer to the other estimators in the calibration estimator family covered by Deville and Särndal (1992), as opposed to the GREG estimator.

Although almost all surveys in practice are subject to frame deficiencies and nonresponse, the theories in Deville and Särndal (1992) are developed for the ideal situation where non-sampling errors do not exist. In the purely sampling context, many alternative forms of calibration weighting are asymptotically identical. This leads to a breakthrough in our understanding of some commonly used calibration estimators that do not have closed-form

solutions, such as raking. As a result, the GREG estimator is often considered a good approximation of the general calibration estimators. However, non-sampling errors such as nonresponse almost always exist in real-world surveys. In the past decade, Särndal and Lundström (1999, 2005), Kott and Chang (2006, 2008, 2010), and Thibaudeau and Slud (2009) have proposed different methods for using calibration to correct nonresponse bias through one-step weighting, yet we still lack understanding of the empirical properties of the calibration estimators generated by these methods. For example, it is unclear whether (or under what conditions) the GREG estimator and the general calibration estimators are asymptotically equivalent when nonresponse is present in a survey and calibration weighting is used to reduce potential nonresponse bias. In practice, poststratification and raking are both widely used in the US and European surveys, and the corresponding estimators are the special cases of the GREG estimator and the general calibration estimator, respectively. Quite often survey practitioners choose between the two estimators based on the availability of the benchmark totals and the case counts in the survey requiring calibration with the hope that poststratification and raking reduce potential nonresponse bias to a similar extent and thus result in "approximately equivalent" estimators. Sometimes only the marginal totals are available, and the practitioner may have no choice but to use raking, even if poststratification might have done a better job in reducing nonresponse bias. No systematic research has been conducted on comparing the performance of the poststratification estimator and the raking estimator when calibration is used to correct nonresponse bias.

The first contribution we attempt to make to the literature is to release the assumption of no non-sampling error and evaluate the properties of the calibration estimators when calibration is used to correct nonresponse bias through a one-step weighting approach. In the absence of non-sampling errors, the purely design-based properties of the calibration estimators were assessed by Deville and Särndal (1992). When nonresponse exists, however, the properties of a calibration estimator may depend on the underlying outcome variable model and response model. Our research evaluates the impacts of sampling, population structure, and response mechanism simultaneously through analytical work and simulation studies. The theoretical and empirical results provide survey practitioners with guidance of how to evaluate different calibration estimators and choose between them under various population structure models and response mechanism models.

The rest of the article is organized as follows: Section 2 summarizes the literature on the properties of various calibration estimators in the purely sampling context as well as the research on using calibration for nonresponse adjustment through single-step weighting. Sections 3 and 4 attempt to fill in some gaps in the literature through analytical work and a simulation study, respectively. Section 5 summarizes the findings briefly and discusses the direction of our future research.

2. Literature Review

2.1 Calibration in Absence of Non-sampling Error

There are two different approaches for incorporating auxiliary information in the estimation, labeled "regression approach" and "calibration approach". Under the umbrella of calibration approach, Section 2.1.1 explains how to use the distance function method to obtain a calibration estimator, followed by the description of an alternative method called the function form method in Section 2.1.2. Section 2.1.3 presents the existing theories developed by Deville and Särndal (1992) on the comparison of the GREG estimator and the general calibration estimator. It is important to note that all the theory presented in this section was developed for the situations absent of non-sampling error.

2.1.1 Two Approaches to Incorporate Auxiliary Information in Estimation

There are two systematic ways to take account of auxiliary information in the estimation. In their original definition of the calibration estimator, Deville and Särndal (1992) require "minimum distance" between the calibration weights and the original sampling weights, subject to satisfying the calibration equation. In general, the term "calibration approach" often refers to creating estimators by benchmarking the auxiliary information to external controls.

For a sample s drawn from a population U, let y_k be the value of the variable of interest, y, for the kth population element, which is associated with an auxiliary vector value, $\mathbf{x}_k = (x_{k1}, \ldots x_{kp}, \ldots, x_{kP})^{\mathrm{T}}$. For the elements $k \in s$, we

observe $(y_k, \mathbf{x}_k)$. For simplicity, the population total of $\mathbf{x}$, $\mathbf{t}_x = \sum_U \mathbf{x}_k$, which is often referred to as the benchmark control vector, is assumed to be accurately known.

The objective is to estimate the population total $t_y = \sum_U y_k$. Let d_k be the basic sampling design weight calculated as the inverse of the inclusion probability π_k. The Horvitz-Thompson estimator is $\hat{t}_{y\pi} = \sum_s y_k / \pi_k = \sum_s d_k y_k$. The calibration estimator is defined as $\hat{t}_{yw} = \sum_s w_k y_k$, with weights w_k as close as possible, in an average sense based on a distance function, to the basic design weights d_k while respecting the calibration equation

$$\sum_s w_k \mathbf{x}_k = \mathbf{t}_x \tag{2.1}$$

Under a chosen distance function $G_k(w_k, d_k)$, this becomes an optimization problem. The goal is to find a set of weights $\{w_k\}_{k\in s}$ that minimizes $\sum_{k\in s} G_k(w_k, d_k)$ subject to (2.1). This leads to the Lagrange function

$$\Psi = \sum_{k\in s} G\left(w_k, d_k\right) + \boldsymbol{\lambda}^{\mathrm{T}} \left(\mathbf{t}_x - \sum_{k\in s} w_k \mathbf{x}_k\right) \tag{2.2}$$

which is minimized to find the optimal set of weights $\{w_k\}_{k\in s}$.

The calibration weights can be expressed as

$$w_k = d_k F_k(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}) \tag{2.3}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_1, \ldots, \lambda_P)^{\mathrm{T}}$ is the vector of Lagrange multipliers determined from (2.2). $\boldsymbol{\lambda}$ corresponds to a realized sample, but for simplicity we sometimes use $\boldsymbol{\lambda}$ as the shorthand for $\boldsymbol{\lambda}_s$. $F_k(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda})$ is the inverse function of $g_k(w_k, d_k) = \partial G_k(w_k, d_k) / \partial w_k$, the first derivative of the distance function taken with respect to the calibration weight. $F_k\left(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}\right)$ uniquely corresponds to $G_k(w_k, d_k)$. It is assumed that F_k is non-negative and convex, and that $F_k(1) = 0$, implying that when $w_k = d_k$ the distance between the basic design weights and calibrated weights is zero. Moreover, it is required that $F_k^{'}$ is continuous, monotonic, and that $F_k^{'}(1) = 0$ and $F_k^{''}(1) > 0$, which makes $w_k = d_k$ a local minimum.

The Horvitz-Thompson estimator of $\mathbf{t}_x$ is $\hat{\mathbf{t}}_{x\pi} = \sum_s d_k \mathbf{x}_k$, so the calibration equation can be expressed as:

$$\sum_s d_k F_k(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}) \mathbf{x}_k - \sum_s d_k \mathbf{x}_k = \mathbf{t}_x - \hat{\mathbf{t}}_{x\pi} \tag{2.4}$$

Define

$$\Phi_s\left(\boldsymbol{\lambda}\right) = \sum_s d_k \left\{F_k\left(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}\right) - 1\right\} \mathbf{x}_k \tag{2.5}$$

Then (2.4) can be written as

$$\Phi_s\left(\boldsymbol{\lambda}\right) = \mathbf{t}_x - \hat{\mathbf{t}}_{x\pi} \tag{2.6}$$

The task of obtaining w_k boils down to solving (2.6) for $\boldsymbol{\lambda}$. The calibration estimator of t_y is

$$\hat{t}_{yw} = \sum_s w_k y_k = \sum_s d_k F_k(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}) y_k \tag{2.7}$$

Depending on the distance function $G_k(w_k, d_k)$, an iterative process may be required to obtain a solution for $\boldsymbol{\lambda}$. There is no non-sampling error in Deville and Särndal's setup, so the Horvitz-Thompson estimator $\hat{t}_{y\pi}$ using basic sampling weights d_k is an unbiased estimator of the true population total. If the calibration weights w_k are as close as possible, according to $G_k(w_k, d_k)$, to the basic sampling weights d_k, then a realistic expectation is that the calibration weights will maintain near unbiasedness.

Although several distance functions are discussed in Deville and Särndal (1992), most theoretical research has focused on the GLS distance function $\sum_s (w_k - d_k)^2 / d_k q_k$, where $1/q_k$ is the positive weight associated with the *k*th term and is unrelated to d_k. Under this distance function, the calibration equation has a closed-form solution. We obtain $\mathrm{F}_k(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}) = 1 + q_k \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}$, and the calibration estimator is the GREG estimator

$$\hat{t}_{yreg} = \sum_s d_k (1 + q_k \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}) y_k = \hat{t}_{y\pi} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})^{\mathrm{T}} \hat{\mathbf{B}}_s \tag{2.8}$$

where

$$\boldsymbol{\lambda} = \mathbf{T}_s^{-1}(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}) \tag{2.9}$$

$$\hat{\mathbf{B}}_s = \mathbf{T}_s^{-1} \sum_s d_k q_k \mathbf{x}_k y_k \tag{2.10}$$

$$\mathbf{T}_s = \sum_s d_k q_k \mathbf{x}_k \mathbf{x}_k^{\mathrm{T}} \tag{2.11}$$

An alternative method for obtaining the calibration estimator is referred to as the "regression approach". With the regression approach, estimators are calculated by using an assisting model that closely represents the relationship between the outcome variable and the auxiliary variables. The assisting model is also referred to as the calibration model or the working prediction model by Kott (2006) to distinguish it from other models such as those used to address response propensity. The assisting model can have linear or nonlinear forms. When the assisting model is a linear regression model, the weight happens to be calibrated to the auxiliary controls and the estimator (which is the GREG estimator) is expressible as a linearly weighted sum with calibrated weights as a by-product. One advantage of the GREG estimator is that the calibrated weights are independent of any particular outcome variable y and can therefore be applied to all the variables of interest in a survey.

Our research adopts the perspectives of both approaches. The weights are primarily justified by their consistency with the benchmark controls (which is the calibration approach). Although the calibration approach does not refer explicitly to any assisting models, we demonstrate that the performance of a calibration estimator in the presence of nonresponse depends on the choice of auxiliary vector and/or function form used in the calibration process, and this requires a modeling effort in some sense.

2.1.2 Distance Function Method versus Function Form Method

Under the umbrella of the calibration approach, two methods are discussed in the literature. Deville and Särndal (1992) initially require that the set of calibration weights $\{w_k\}_{k \in s}$ minimize some distance function $\sum_{k \in s} G_k(w_k, d_k)$ subject to satisfying the calibration equation – This is the "distance function method" described in Section 1.1. An alternative approach is the "function form method" or "instrumental vector method" (Estevao and Särndal 2006, Kott 2006), which can also generate many alternative sets of weights calibrated to the same auxiliary information.

The function form method removes the limitation that the calibration weights minimize a distance function, and requires only that $\{w_k\}_{k \in s}$ satisfy the calibration equation and be of the function form $w_k = d_k F(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\lambda})$, where d_k is the design weight, and $\mathbf{z}_k$ is a vector with values defined for the units in the sample and sharing the dimension of the specified benchmark control vector $\mathbf{x}_k$. The vector $\mathbf{z}_k$ is called the instrumental vector for the calibration, which can be a specified function of $\mathbf{x}_k$ or of other background data about unit *k* (Särndal and Lundström 2005). The vector $\boldsymbol{\lambda}$ is determined from the calibration equation. The function $F(\cdot)$ plays a similar role as $G_k(w_k, d_k)$ does in the distance minimization method. For easy reference, we refer to $F(\cdot)$ as "weight adjustment function" or "adjustment function" in our research. One possible form of the weight adjustment function is $w_k = d_k(1 + \mathbf{z}_k^{\mathrm{T}}\boldsymbol{\lambda})$, and the corresponding calibration estimator is:

$$\hat{t}_{ycal} = \sum_s d_k (1 + \mathbf{z}_k^{\mathrm{T}}\boldsymbol{\lambda}) y_k \tag{2.12}$$

where

$$\boldsymbol{\lambda} = (\sum_s d_k \mathbf{x}_k \mathbf{z}_k^{\mathrm{T}})^{-1}(\mathbf{t}_x - \sum_s d_k \mathbf{x}_k) \tag{2.13}$$

The GREG estimator $\hat{t}_{yreg}$ defined in (2.8) is a special case of (2.12) obtained for $\mathbf{z}_k = q_k \mathbf{x}_k$.

When nonresponse exists in a survey, we think that it is more appropriate to understand the calibration process using the function form method rather than the distance function method. This is because in the presence of nonresponse, the Horvitz-Thompson estimator for the total of an outcome variable y using the basic design weights becomes $\hat{t}_{y\pi} = \sum_r d_k y_k$, where r represents the responding set. This estimator is biased when $r \neq s$. If the calibration process aims to correct the nonresponse bias, it is neither necessary nor appropriate to require the calibrated weights to be "as close as possible" to the basic design weights based on a distance function.

More discussions about the weighting adjustment function $F(\cdot)$ are included in Section 2.2.1. When applying the function form method in practice, survey practitioners face questions such as how to choose the variables to be included in the vector $\mathbf{z}_k$, or what is the advantage of making the form of $\mathbf{z}_k$ different from the calibration vector $\mathbf{x}_k$. These questions have not been clearly answered by the existing literature. Särndal (2007) gives an example showing that "even 'deliberately awkward choices' for $\mathbf{z}_k$ give surprisingly good results". However, the conclusion of the near-unbiasedness of the calibration estimator in this situation seems to depend on the assumption of no non-sampling error, which may not hold in the presence of nonresponse.

2.1.3 Relationship between GREG Estimator and General Calibration Estimators

As described in Section 1.1, various calibration estimators can be derived with the aid of different distance measures under the same set of constraints on the auxiliary variables. Alternative distance functions are compared in Deville, Särndal, and Sautory (1993), Singh and Mohl (1996), and Stukel, Hidiroglou, and Särndal (1996). In a purely sampling context, there are usually very small differences between the point estimates corresponding to the various distance functions, and changes in the distance function often have minor effects on the variance of the calibration estimator even if the sample size is rather small. In particular, the GREG estimator and the other members of the calibration estimator family (referred to as the "general calibration estimators") are compared in Deville and Särndal (1992). They conclude that the GREG estimator is a first approximation to the general calibration estimators; all the general calibration estimators are asymptotically equivalent to the GREG; and the variance estimator for the GREG can be used for the general calibration estimators. Although the GREG estimator is a special case of the calibration estimator family when the function form is $F(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}) = 1 + q_k \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}$, we use $\hat{t}_{yreg}$ to denote the GREG estimator and $\hat{t}_{yw}$ to denote the other calibration estimators (i.e., the general calibration estimators) for the purpose of comparison.

Deville and Särndal (1992) consider a sequence of finite populations and sampling designs indexed by *n*, where *n* is the sample size (for a fixed-sized sampling design) or the expected sample size (for a random-sized sample design). The finite population size, *N*, tends to infinity with *n*. Several assumptions are made about the auxiliary vector **x**: (i) $\lim N^{-1}\mathbf{t}_x$ exists; (ii) $N^{-1}\left(\hat{\mathbf{t}}_{x\pi} - \mathbf{t}_x\right) = O_p(n^{-1/2})$, where the subscript p means probability induced by the sample; and (iii) $n^{1/2}N^{-1}(\hat{\mathbf{t}}_{x\pi} - \mathbf{t}_x)$ converges in distribution to the multinormal $N(\mathbf{0}, \mathbf{A})$. Two additional assumptions are also made for proving their Results 3-5: (iv) $\max\|\mathbf{x}_k\| = M < \infty$, where max is over *n* as well as over *k*; and (v) $\max F_k''(0) = M' < \infty$. Assumptions (i) through (iii) have two practical implications. First, the components of $\hat{\mathbf{t}}_{x\pi} - \mathbf{t}_x$ are considered small and quantities on the order of $\left\|\hat{\mathbf{t}}_{x\pi} - \mathbf{t}_x\right\|^2$ are considered negligible. Second, $\hat{\mathbf{t}}_{x\pi} - \mathbf{t}_x$ follows an approximately normal distribution with covariance matrix $n^{-1}N^2\mathbf{A}$ (where **A** can be viewed as a matrix that describes an asymptotic effect of the sampling design used for the survey), and this is to justify the use of the normal approximation in confidence intervals based on $\hat{t}_{yw}$. Assumption (iv) is usually satisfied in practice. Assumption (v) is verified for all the calibration estimators given in Deville and Särndal (1992).

Deville and Särndal (1992) show five results. Result 1 states that the calibration equation (2.6) has a unique solution belonging to an open neighborhood of **0**, with probability tending to 1 as $n \to \infty$. Results 2 and 3 are about the

magnitude of the Lagrange multiplier. They prove that $\boldsymbol{\lambda}_s = \mathbf{T}_s^{-1}\left(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}\right) + O_p\left(n^{-1}\right) = O_p(n^{-1/2})$, so $\boldsymbol{\lambda}_s$ tends to $\mathbf{0}$ in design probability as $n \to \infty$. Result 4 indicates that the general calibration estimators are design-consistent, and the difference between the general calibration estimators and the Horvitz-Thompson estimator is asymptotically zero. That is, $N^{-1}\left(\hat{t}_{yw} - \hat{t}_{y\pi}\right) = O_p(n^{-1/2})$. Result 5 compares the general calibration estimators with the GREG estimator. For any weight adjustment function $F_k(\cdot)$ obeying their assumptions, $\hat{t}_{yw}$ given by equation (2.7) is asymptotically equivalent to the GREG estimator given by equation (2.8), in the sense that $N^{-1}\left(\hat{t}_{yw} - \hat{t}_{yreg}\right) = O_p\left(n^{-1}\right)$. Results 4 and 5 together show that as $n \to \infty$, the difference between the general calibration estimators and the GREG estimator approaches to zero faster than the difference between the general calibration estimators and the Horvitz-Thompson estimator. The asymptotic variance of $\hat{t}_{yw}$ is thus the same as that of the GREG estimator.

The results above have important practical implications because many general calibration estimators do not have a closed-form solution. For example, although the raking ratio method has a long history, the variance of the raking estimator is difficult to derive even approximately. Deville and Särndal (1992) resolve the problem by using the property that the general calibration estimators and the GREG estimator are asymptotically equivalent. Thus the large-sample variance of the raking ratio estimator can be calculated using the same formula as that for the GREG estimator, given in Särndal, Swensson, and Wretman (1992).

It is important to note that all the results in Deville and Särndal (1992) are derived under the assumptions i) $N^{-1}\left(\hat{\mathbf{t}}_{x\pi} - \mathbf{t}_x\right) = O_p(n^{-1/2})$; and ii) $n^{1/2}N^{-1}(\hat{\mathbf{t}}_{x\pi} - \mathbf{t}_x)$ converges in distribution to a multinormal distribution with mean of $\mathbf{0}$. That is, they require the Horvitz-Thompson estimator with the basic design weights to be unbiased and consistent, which is true only in the purely sampling context. When non-sampling errors exist, it is unclear whether the GREG estimator is still asymptotically equivalent to other calibration estimators.

2.2 Calibration for Nonresponse Bias Reduction

Our research focuses on the non-sampling error caused by nonresponse. To keep the picture simple, we assume that the sampling frame has perfect coverage and there is no measurement error in the survey. In practice, calibration is widely used to correct nonresponse bias in government-sponsored studies such as the National Health and Nutrition Examination Survey and Medical Expenditure Panel Survey. There are several variations in the literature on how to adjust for nonresponse and calibrate the weights to benchmark controls. The conventional approach uses auxiliary information in two steps (Kalton and Flores-Cervantes 2003). In step (i), a response model is formed based on the patterns of correlation between the response probabilities and available auxiliary variables. The aim is to derive good proxies of the unknown response probabilities, so as to limit the nonresponse bias as much as possible. In step (ii), the goal is to select the auxiliary variables that best meet the dual purpose of reducing the sampling variance and of giving added protection against nonresponse bias. The conventional approach is embodied in the estimator $\hat{t}_y = \sum_r d_k(1/\hat{p}_k)y_k$, where d_k is the design weight calculated as the inverse of the selection probability π_k , $\hat{p}_k$ is the estimated response propensity, and r is the set of respondents. Survey practitioners usually act (for the purpose of variance estimation, for example) as if $\pi_k\hat{p}_k$ was the true selection probability of element k . An unavoidable bias results from replacing the unknown p_k with $\hat{p}_k$ based on limited auxiliary information. An alternative approach is to skip explicitly estimating the response propensity, but use calibration for nonresponse adjustment directly. This approach has the potential to simplify the derivation of the variance estimation formulas, so we adopt this approach in our work. A single-step weighting approach through calibration was first proposed by Särndal and Lundström (1999, 2005). The literature has been expanded by Kott (2005) and Chang and Kott (2008, 2010) in the past decade.

2.2.1 Särndal and Lundström Method

In the Särndal and Lundström method, all the auxiliary controls, from either the population (with a control vector of dimension J^*) or the sample (with a control vector of dimension J^o), are included in the calibration equation, with the dual purpose of reducing both sampling error and nonresponse bias. The auxiliary vector $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^o \end{pmatrix}$ has dimension

$J^* + J^o$. The corresponding information input is $\mathbf{t}_x = \begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_s d_k \mathbf{x}_k^o \end{pmatrix}$. We seek a weighting system w_k for $k \in r$, the respondent set, that satisfies the calibration equation $\sum_r w_k \mathbf{x}_k = \mathbf{t}_x$. The calibrated weights are $w_k = d_k v_k$, where v_k corresponds to the weighting adjustment function $F(\cdot)$ described in Section 2.1.2 and can take different forms.

Although the distance function method is used in Lundström and Särndal (1999) for obtaining w_k, their later work adopts the function form method, which seems more appropriate when nonresponse is present and calibration is used to correct nonresponse bias. The calibration equation poses only weak constraints on the weights. Depending on the form v_k takes, there exist many sets of calibrated weights for a given auxiliary vector $\mathbf{x}_k$. Särndal and Lundström (2005) discuss two alternative schemes for defining the function form for v_k: (i) as a function of the auxiliary vector $\mathbf{x}_k$; and (ii) as a function of any vector $\mathbf{z}_k$ (referred to as instrumental vector) specified for $k \in r$ and with the same dimension as $\mathbf{x}_k$. Under scheme (i), v_k should reflect the known individual characteristics of the element $k \in r$, summarized by the vector value $\mathbf{x}_k$. The calibration equation can be expressed as $\sum_r d_k F(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r) = \mathbf{t}_x$, where $\boldsymbol{\lambda}_r$ is a vector to be determined through the calibration equation. A simple function form is recommended that depends linearly on $\mathbf{x}_k$: $F(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r) = 1 + \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r$, where $\boldsymbol{\lambda}_r = (\sum_r d_k \mathbf{x}_k \mathbf{x}_k^{\mathrm{T}})^{-1} (\mathbf{t}_x - \sum_r d_k \mathbf{x}_k)$. Although nonlinear forms can also be considered such as $F(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r) = \exp(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r)$, Särndal and Lundström (2005) suggest that the linear form will suffice due to its considerable computational advantage and fits the routine production environment. However, little theoretical or empirical justification is provided to support this statement.

An alternative scheme is to define the weighting adjustment function using the instrumental vector $\mathbf{z}_k$, specified for $k \in r$ and with the same dimension as $\mathbf{x}_k$. The vector $\mathbf{z}_k$ can be a specified function of $\mathbf{x}_k$ or any background data about k. Only linear function form based on $\mathbf{z}_k$ is considered by Särndal and Lundström (2005). The calibrated weights are $w_k = d_k (1 + \mathbf{z}_k^{\mathrm{T}} \boldsymbol{\lambda}_r)$, where $\boldsymbol{\lambda}_r = (\sum_r d_k \mathbf{x}_k \mathbf{z}_k^{\mathrm{T}})^{-1} (\mathbf{t}_x - \sum_r d_k \mathbf{x}_k)$. Besides alerting the reader to this generality of the calibration approach, Särndal and Lundström (2005) give little information about how to choose $\mathbf{z}_k$ except to suggest that $\mathbf{z}_k = \mathbf{x}_k$ is the "standard choice".

Särndal and Lundström (2005) claim that their approach meets the double objective of reducing sampling error and nonresponse error in the presence of powerful auxiliary information. A variance estimator is also developed to take into account the increased variance caused by nonresponse, as described in Section 2.2.

We can see that scheme (ii) is the generalization of scheme (i) in Särndal and Lundström (2005). When $\mathbf{z}_k = \mathbf{x}_k$, the two schemes give identical estimators. Furthermore, when $r = s$ (indicating full response) and $\mathbf{x}_k = \mathbf{x}_k^*$ (meaning that the auxiliary vector contains information only from external benchmarks and not from the sampling frame), the calibration estimator $w_k = d_k (1 + \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r)$ and the GREG estimator defined in equation (2.8) are identical.

2.2.2 Kott and Chang Method

Recent development by Kott (2006) and Chang and Kott (2008, 2010) emphasize that the set of variables modeling the response mechanism (referred to as "model variables") can be different from the benchmark variables in the calibration equation. The vector for the benchmark controls in the calibration equation is still $\mathbf{x}_k$, with known population totals $\mathbf{t}_x$. Unit nonresponse is viewed as an additional phase of Poisson sampling. Using the quasi-randomization perspective, each element k in the original sample is assumed to have a response probability $p_k(\cdot)$, which is a function of the response model covariate vector $\mathbf{z}_k$. The statistician can specify the function form for $p_k(\cdot)$ and the unknown parameters in the function can be estimated. Kott (2006) explains why it may be desirable to use a vector $\mathbf{z}_k$ that is different from $\mathbf{x}_k$ – Sometimes the variables the response mechanism depends on are known

only for respondents, not for the whole sample. For example, in an agriculture survey, the benchmark variables can be previous-census frame variables known for every farm in the population while the model variables are current-period variables known only for survey respondents. Kott (2006) still requires that the dimensions of $\mathbf{z}_k$ and $\mathbf{x}_k$ coincide. Chang and Kott (2008) expand the method so that it allows the number of benchmark variables (i.e., the dimension of $\mathbf{x}_k$) to exceed the number of response model covariates (i.e., the dimension of $\mathbf{z}_k$).

Based on Kott and Chang, the response propensity for each responding unit k can be specified as $p(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\beta})$, an unknown but estimable linear combination of the response model covariate vector $\mathbf{z}_k$. The input weight for the calibration equation is calculated as the product of basic design weight d_k and $1/p(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\beta})$, and then the vector $\boldsymbol{\beta}$ can be estimated from the data using the calibration equation $\mathbf{t}_x = \sum_{k\in r}\frac{d_k}{p(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\beta})}\mathbf{x}_k$. This equation is sufficient to determine $\hat{\boldsymbol{\beta}}$ if the dimension of $\mathbf{x}_k$ equals the dimension of $\mathbf{z}_k$ (Kott 2006). On the other hand, when the dimension of $\mathbf{x}_k$ exceeds the dimension of $\mathbf{z}_k$, the calibration equation can be modified into a nonlinear regression-type equation $\mathbf{t}_x = \sum_{k\in r}\frac{d_k}{p(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\beta})}\mathbf{x}_k + \boldsymbol{\varepsilon}$, where $\mathbf{z}_k$ and $\mathbf{x}_k$ denote the vectors for response model covariates and benchmark variables respectively, $\mathbf{t}_x$ is the calibration target values consisting the known population totals, and $\boldsymbol{\varepsilon}$ is the error term between the calibrated estimates and the population controls of the auxiliary variable (Chang and Kott 2008). Although in theory, the response propensity $p(\cdot)$ can take different forms, the discussions in Chang and Kott (2008) are restricted to the situation where the response propensity is the linear function of the response model covariates.

2.3 Summary of Gaps in the Literature

In Sections 2.1 and 2.2, we provided a review of the calibration literature. In this section, we note gaps in the literature that remain to be addressed.

First, there is little research on evaluating the asymptotic properties of different calibration estimators when nonresponse is present and calibration is used for correcting nonresponse bias. For example, the raking ratio estimator and poststratification estimator are special cases of the general calibration estimator and GREG estimator respectively and widely used in practice. Based on Deville and Särndal (1992), these two estimators are asymptotically equivalent in absence of non-sampling errors. However, non-sampling errors such as nonresponse error almost always exist in surveys. It is important to re-examine conclusions in Deville and Särndal (1992) in the context of using calibration for nonresponse adjustment.

Second, if the conclusions in Deville and Särndal (1992) do not hold when calibration is used for nonresponse adjustment, then the existing literature provides neither a good framework for comparing the performances of different calibration estimators, nor practical guidance for choosing the appropriate auxiliary vectors and/or function forms for calibration weighting. To answer these questions, we need to go beyond the purely design-based approach used in Deville and Särndal (1992) and examine the underlying models for population structure (i.e., what variables are correlated with the key outcome variable) and response mechanism (i.e., what variables are corrected with response). Survey practice calls for guidelines for how to select variables to be included in the auxiliary vectors $\mathbf{x}_k$ and response model covariate vector $\mathbf{z}_k$ when calibration is used for nonresponse adjustment, but there is very little research in this area.

3. Comparison of the Design-Based Properties of GREG Estimator and General Calibration Estimator in the Presence of Nonresponse

This section attempts to fill in one gap in the literature by comparing the asymptotic properties of the general calibration estimator and GREG estimator when calibration is used for nonresponse adjustment through a single-step weighting approach. We use the Särndal and Lundström calibration method described in Section 2.2.1 and focus on the situation where the instrumental vector $\mathbf{z}_k$ coincides with the calibration vector $\mathbf{x}_k$. In the presence of nonresponse, the Horvitz-Thompson estimator of the total for the auxiliary vector using the basic design weights is a

function of the respondent set and can therefore be "far" from the benchmark control total. This violates one of the key assumptions in Deville and Särndal (1992), so it is unclear whether their conclusions still hold.

The setup and analytical work in this section largely follow the approach taken by Deville and Särndal (1992), which is purely design-based. We use the terms "new assumption" and "new result" to differentiate our assumptions and findings from those in Deville and Särndal (1992). Our theoretical results are applicable to a family of general calibration estimators. At the end of this section, we point out the limitations of the purely design-based approach and emphasize the importance of examining the underlying models for the outcome variable and response rate when comparing different calibration estimators.

3.1 Scope and Assumptions

First, we assume the analytic survey (i.e., the survey requiring calibration) and benchmark survey come from the same population U of size N. Although the benchmark control totals are often estimated and subject to sampling and non-sampling errors in practice, we assume that the total for the auxiliary vector $\mathbf{x}$ is known and equal to the true population total $\mathbf{t}_x = \sum_U \mathbf{x}_k$.

Second, we assume that the analytic survey has no coverage or measurement error, but may suffer from nonresponse error that can bias the estimates of parameters such as population totals. In the presence of nonresponse, the survey has a respondent set r of size n_r. We assume that no separate nonresponse adjustment is conducted prior to calibration, so that in the absence of calibration, the population estimates are calculated using only the basic design weights d_k, i.e., using the Horvitz-Thompson estimators of the population totals of the auxiliary vector and outcome variable $\hat{\mathbf{t}}_{r_{x\pi}} = \sum_r d_k \mathbf{x}_k$ and $\hat{t}_{r_{y\pi}} = \sum_r d_k y_k$ respectively. The proofs in this section require no explicit specifications for the sample design of the analytic survey, but it is reasonable to assume that the units in the survey are selected with a method that results in unbiased estimates of the totals for various variables in the absence of nonresponse.

Finally, although survey nonresponse is caused by a random mechanism, for the simplicity of theoretical derivation, we assume that each population member has fixed response propensity of either 1 or 0. That is, a population member either always agrees or always refuses to participate in the analytic survey. In the presence of nonresponse, the design-based expectation of the Horvitz-Thompson estimator reflects the characteristics of the "responding population" U_r of size N_r. We define $E_\pi(\hat{\mathbf{t}}_{r_{x\pi}}) = \mathbf{t}_{r_{x\pi}}$ and $E_\pi(\hat{t}_{r_{y\pi}}) = t_{r_{y\pi}}$, where E_π means design-based expectation, and $\mathbf{t}_{r_x}$ and $\mathbf{t}_{r_y}$ are the totals of the auxiliary variables and outcome variable for the respondent population U_r, respectively. It is reasonable to assume that the size of the responding population, N_r, is large so that $N_r/N = O(1)$.

The theoretical derivation in this section requires the following assumptions. We refer to these as "new assumptions" in contrast of those in Deville and Särndal (1992).

New assumption (i): $\lim N_r^{-1}\mathbf{t}_{r_x}$ exist, but in general, $\lim N_r^{-1}\mathbf{t}_{r_x} \neq N^{-1}\mathbf{t}_x$.

New assumption (ii): $N_r^{-1}(\hat{\mathbf{t}}_{r_{x\pi}} - \mathbf{t}_{r_x}) \to \mathbf{0}$ in design probability. $N_r^{-1}\left(\hat{\mathbf{t}}_{r_{x\pi}} - \mathbf{t}_{r_x}\right) = O_p(n_r^{-1/2})$.

New assumption (iii). $n_r^{1/2}N_r^{-1}(\hat{\mathbf{t}}_{r_{x\pi}} - \mathbf{t}_{r_x})$ converges in distribution to the multinormal $N(\mathbf{0}, \mathbf{A})$, where $\mathbf{A}$ can be viewed as a matrix that describes an asymptotic effect of the sampling design used for the analytic survey.

Recall that one of the key assumptions in Deville and Särndal (1992) is that in the purely sampling context, the Horvitz-Thompson estimators of the population totals of the auxiliary vector approach to the true values of the population as the sample size increases. That is, $N^{-1}\left(\hat{\mathbf{t}}_{x\pi} - \mathbf{t}_x\right) = O_p(n^{-1/2})$. Based on our new assumption (ii), the Horvitz-Thompson estimators from the respondent set approach only to $\mathbf{t}_{r_x} = E_\pi(\hat{\mathbf{t}}_{r_{x\pi}})$. We know that $\mathbf{t}_{r_x} \neq \mathbf{t}_x$ in the presence of nonresponse. This has important implications in the theoretical derivation below.

3.2 Analytical Results

In this section we re-examine the results in Deville and Särndal (1992) in the context of using calibration for nonresponse adjustment through single-step weighting. The input weights for the calibration equation are the basic design weights d_k. In this setup, the Horvitz-Thompson estimator $\hat{t}_{r_{y\pi}}$ using the basic sampling weights d_k is biased due to nonresponse (unless the nonresponse is generated by a missing completely at random mechanism), so calibration is used to reduce such bias to the extent possible. It is appropriate to conduct calibration using the function form method instead of the distance function method, and the calibration weights may not be as "close" to the basic design weights as required in Deville and Särndal (1992). We suspect that whether the calibration equation has a solution may depend on the overall response rate as well as how the response rates differ by subgroups formed by the variables used as benchmark controls in the calibration. We show that the vector for the Lagrange multiplier determined from the calibration equation, $\boldsymbol{\lambda}_r$, consists of a term that is driven by the difference between the Horvitz-Thompson estimator of the auxiliary vector (using the basic design weights) for the respondent population total (denoted by $\hat{\mathbf{t}}_{r_{x\pi}}$) and the benchmark control total (denoted by $\mathbf{t}_x$). Unless nonresponse is negligible, this term does not decrease as the survey sample size increases, so $\boldsymbol{\lambda}_r$ may tend to a non-zero constant vector in design probability. Our analytical work results in the formulas for: (1) the difference between a general calibration estimator and Horvitz-Thompson estimator; and (2) the difference between a general calibration estimator and the GREG estimator, in the presence of nonresponse. We prove that when nonresponse exists and calibration is used to reduce nonresponse bias through single-step weighting, the general calibration estimators and the GREG estimator are not asymptotically equivalent in general situations.

In the presence of nonresponse, the calibration equation is $\sum_r w_k \mathbf{x}_k = \mathbf{t}_x$ and the calibration estimator is $\hat{t}_{yw} = \sum_r w_k y_k$. Equations (2.5) and (2.6) in Section 2 should be modified into

$$\Phi_r(\boldsymbol{\lambda}) = \sum_r d_k \left\{ F_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right) - 1 \right\} \mathbf{x}_k \tag{3.1}$$

$$\Phi_r(\boldsymbol{\lambda}) = \mathbf{t}_x - \hat{\mathbf{t}}_{r_{x\pi}} = (\mathbf{t}_x - \mathbf{t}_{r_x}) + (\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}) \tag{3.2}$$

Note that in this section, we sometimes use $\boldsymbol{\lambda}$ as a short-handed form of $\boldsymbol{\lambda}_r$ for convenience.

We know that $\mathbf{t}_{r_x} \neq \mathbf{t}_x$ in the presence of nonresponse, so the right-hand side of (3.2) contains a non-zero term that does not exist in equation (2.6) of Section 2. This non-zero term plays an important role in the discussions below. We have five new results in parallel to the ones in Deville and Särndal (1992).

New Result 1. As $n_r \to \infty$, whether equation (3.2) has a solution may depend on the difference between $\mathbf{t}_{r_x}$ and $\mathbf{t}_x$ as well as the function form $F_k(\cdot)$ used in the calibration.

For this result, we give intuitive explanations instead of strict proof. In the presence of nonresponse, the calibration equation can be written as:

$$N_r^{-1}\Phi_r(\boldsymbol{\lambda}) = N_r^{-1}\sum_r d_k F_k(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda})\mathbf{x}_k - N_r^{-1}\sum_r d_k \mathbf{x}_k = N_r^{-1}(\mathbf{t}_x - \mathbf{t}_{r_x}) + N_r^{-1}(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}) \tag{3.3}$$

For the right-hand side of (3.3), the second term is similar to that in Deville and Särndal (1992). $N_r^{-1}(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}) = O_p(n_r^{-1/2})$ and is asymptotically $\mathbf{0}$. However, when nonresponse exists, $\mathbf{t}_x \neq \mathbf{t}_{r_x}$ and the first term $N_r^{-1}(\mathbf{t}_x - \mathbf{t}_{r_x}) = O(1)$. Due to this additional term, the right-hand side of (3.3) does not tend to $\mathbf{0}$, but becomes a non-zero constant vector as n_r increases.

A more intuitive way to understand this result is that in Deville and Särndal (1992), only "small" adjustments need to be made to the basic design weights to obtain the calibration weights, and that is essentially why the calibration

equation almost always has a solution for large samples. When nonresponse exists, the Horvitz-Thompson estimator $\sum_r d_k \mathbf{x}_k$ may be "far" from the benchmark controls $\mathbf{t}_x$ and therefore "large" adjustments on the basic design weights may be required to satisfy the calibration constraints. In this situation, whether the calibration equation has a solution may depend on the difference between $\mathbf{t}_{r_x}$ and $\mathbf{t}_x$ as well as the function form $F_k(\cdot)$ used in the calibration. An empirical example is that for the same calibration constraints and respondent set r, poststratification always has a solution but raking does not always converge.

New Result 2. Let $\boldsymbol{\lambda}_r$ be the solution to equation (3.3) if one exists. If $\mathbf{t}_x - \mathbf{t}_{r_x} \neq \mathbf{0}$, then $\boldsymbol{\lambda}_r = O_p(1)$ in general situations. This means that $\boldsymbol{\lambda}_r$ tends to a non-zero vector in design probability.

Proof: Define $\mathbf{z}_1 = N_r^{-1}(\mathbf{t}_x - \mathbf{t}_{r_x})$ and $\mathbf{z}_2 = N_r^{-1}(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{\pi}})$, so $\boldsymbol{\lambda}_r = (N^{-1}\Phi_r)^{-1}(\mathbf{z}_1 + \mathbf{z}_2)$ if a solution to (3.3) exists. Since $N^{-1}\Phi_r(0) = 0$, we have $\boldsymbol{\lambda}_r - 0 = \left(N^{-1}\Phi_r\right)^{-1}(\mathbf{z}_1 + \mathbf{z}_2) - \left(N^{-1}\Phi_r\right)^{-1}(0)$. Following the notations in Deville and Särndal (1992), we can obtain:

$$\|\boldsymbol{\lambda}_r\| \leq \|\mathbf{z}_1 + \mathbf{z}_2\| K(1-\beta)^{-1} \leq \|\mathbf{z}_1\| K(1-\beta)^{-1} + \|\mathbf{z}_2\| K(1-\beta)^{-1} \tag{3.4}$$

where K is defined in Section 1.1 of Appendix in Deville and Särndal (1992) and $0 < \beta < \frac{1}{2}$.

Since $\mathbf{z}_1 = O(1)$ and $\mathbf{z}_2 = O_p(n_r^{-1/2})$, inequality (3.4) implies that $\boldsymbol{\lambda}_r = O(1) + O_p(n_r^{-1/2})$. The second term tends to $\mathbf{0}$ as n_r increases. However, the first term is a non-zero constant vector in general situations, and does not decrease as n_r increases.

New Result 3. In general situations, $\boldsymbol{\lambda}_r = \mathbf{T}_r^{-1}\left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{\pi}}\right) + O_p(1)$, where $\mathbf{T}_r = \sum_r d_k q_k \mathbf{x}_k \mathbf{x}_k^{\mathrm{T}}$.

Proof: We use $F_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right)$ to denote the adjustment function for a general calibration estimator. For the GREG estimator, the adjustment function takes the form $1 + q_k \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}$. The difference between the two adjustment functions is expressed as:

$$\theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right) = F_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right) - (1 + q_k \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}) \tag{3.5}$$

From (3.1), (3.2) and (3.5), we obtain

$$(\mathbf{t}_x - \mathbf{t}_{r_x}) + (\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{\pi}}) = \sum_r d_k \mathbf{x}_k \left\{ q_k \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r + \theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r\right)\right\} \tag{3.6}$$

Multiplying both sides of (3.6) by $\mathbf{T}_r^{-1}$ and rearranging the terms, we obtain

$$\boldsymbol{\lambda}_r - \mathbf{T}_r^{-1}(\mathbf{t}_x - \mathbf{t}_{r_x}) - \mathbf{T}_r^{-1}(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{\pi}}) = -\mathbf{T}_r^{-1}\sum_r d_k \mathbf{x}_k \theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r\right) \tag{3.7}$$

An important assumption in Deville and Särndal (1992) is that $F_k''(0)$ is uniformly bounded, which is equivalent to $\theta\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r\right) = \max \theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r\right) = O\left(\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r\right)^2\right)$. Note that this assumption requires the condition that $\boldsymbol{\lambda}_r = O_p(n^{-1/2})$, which does not necessarily hold when $\mathbf{t}_x \neq \mathbf{t}_{r_x}$. But given that $\max |\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r| < \infty$, when nonresponse in the analytic survey is not extremely severe, we can still assume that for any $\varepsilon > 0$, there exists K'' such that, for all k, $|\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}| < \varepsilon$ will imply that $\theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right) \leq K''\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right)^2$.

For $\boldsymbol{\lambda}_r$ sufficiently small (which happens when nonresponse in the analytic survey is not extremely severe),

$$\left\|\boldsymbol{\lambda}_r - \mathbf{T}_r^{-1}(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}})\right\| \le \left\|(N_A^{-1}\mathbf{T}_r)^{-1}\right\| K'' \left\{ N_A^{-1}\sum_r d_k \left\|\mathbf{x}_k\right\|^3 \right\} \left\|\boldsymbol{\lambda}_r\right\|^2 + \mathbf{T}_r^{-1}(\mathbf{t}_x - \mathbf{t}_{r_x}) \tag{3.8}$$

We know that $\left\|(N_r^{-1}\mathbf{T}_r)^{-1}\right\| = O_p(1)$ and $N_r^{-1}\sum_r d_k \left\|\mathbf{x}_k\right\|^3 = O_p(1)$. Based on the New Result 2, $\left\|\boldsymbol{\lambda}_r\right\|^2 = O_p(1)$, so the first term of the right-hand side of (3.8) is $O_p(1)$. The second term of the right-hand side of (3.8) is also $O_p(1)$. So we have $\boldsymbol{\lambda}_r = \mathbf{T}_r^{-1}\left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}\right) + O_p(1)$. Although $\mathbf{T}_r^{-1}\left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}\right)$ tends to $\mathbf{0}$ as $n_A \to \infty$, the magnitude of $\boldsymbol{\lambda}_r$ is $O_p(1)$ in general situations. Unless $\mathbf{t}_x = \mathbf{t}_{r_x}$, $\boldsymbol{\lambda}_r$ does not tend to $\mathbf{0}$ as $n_A \to \infty$.

New Result 4. The difference between the general calibration estimator and the Horvitz-Thompson estimator can be expressed in two ways.

In terms of totals:

$$\hat{t}_{r_{yw}} - \hat{t}_{r_{y\pi}} = \widehat{\mathbf{B}}_r^{\mathrm{T}}\left(\mathbf{t}_x - \mathbf{t}_{r_x}\right) + \widehat{\mathbf{B}}_r^{\mathrm{T}}\left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}\right) + (\mathbf{Y}_r - \hat{\mathbf{Y}}_r)^{\mathrm{T}}\mathbf{D}_r\boldsymbol{\theta}_r \tag{3.9}$$

where

$$\underset{p\times 1}{\hat{\mathbf{B}}_r} = \mathbf{T}_r^{-1}\mathbf{X}_r^{\mathrm{T}}\mathbf{D}_r\mathbf{Q}_r\mathbf{Y}_r = \left(\sum_r d_k q_k \mathbf{x}_k \mathbf{x}_k^{\mathrm{T}}\right)^{-1} \sum_r d_k q_k \mathbf{x}_k y_k$$

$$\underset{p\times p}{\mathbf{T}_r} = \mathbf{X}_r^{\mathrm{T}}\mathbf{D}_r\mathbf{Q}_r\mathbf{X}_r = \sum_r d_k q_k \mathbf{x}_k \mathbf{x}_k^{\mathrm{T}}$$

$$\underset{n_r\times p}{\mathbf{X}_r} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n_r 1} & \cdots & x_{n_r p} \end{pmatrix} = (\mathbf{x}_1^{\mathrm{T}}, \ldots \mathbf{x}_k^{\mathrm{T}}, \ldots, \mathbf{x}_{n_r}^{\mathrm{T}})^{\mathrm{T}}$$

$$\underset{n_r\times n_r}{\mathbf{Q}_r} = \begin{pmatrix} q_1 & & 0 \\ & \ddots & \\ 0 & & q_{n_r} \end{pmatrix}$$

$$\underset{n_r\times n_r}{\mathbf{D}_r} = \begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_{n_r} \end{pmatrix}$$

$$\mathbf{Y}_r = (y_1, \ldots y_k, \ldots, y_{n_r})^{\mathrm{T}}$$

$$\hat{\mathbf{Y}}_r = \mathbf{X}_r\hat{\mathbf{B}}_r$$

$$\boldsymbol{\theta}_r = (\theta_1, \ldots \theta_k, \ldots, \theta_{n_r})^{\mathrm{T}}$$

In terms of means:

$$N^{-1}\hat{t}_{r_{yw}} - N_r^{-1}\hat{t}_{r_{y\pi}} = \widehat{\mathbf{B}}_r^{\mathrm{T}}\boldsymbol{\mu}_x\left(1 - (\boldsymbol{\mu}_{r_x}/\boldsymbol{\mu}_x)p\right) - \hat{\boldsymbol{\mu}}_{r_{y\pi}}(1-p) + N^{-1}\widehat{\mathbf{B}}_r^{\mathrm{T}}\left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}\right) + N^{-1}(\mathbf{Y}_r - \hat{\mathbf{Y}}_r)^{\mathrm{T}}\mathbf{D}_r\boldsymbol{\theta}_r \tag{3.10}$$

where $\boldsymbol{\mu}_{r_x}$ is the mean of the auxiliary vector for the respondent population, $\boldsymbol{\mu}_x$ is the true population mean, $\hat{\boldsymbol{\mu}}_{r_{y\pi}}$ is the Horvitz-Thompson estimator of the mean for the outcome variable estimated from the respondent set, and p is the response rate of the analytic survey. In the special situation where $\boldsymbol{\mu}_{r_x} = \boldsymbol{\mu}_x$ (indicating ignorable nonresponse), the difference between the two estimators becomes:

$$N^{-1}\hat{t}_{r_{yw}} - N_r^{-1}\hat{t}_{r_{y\pi}} = (\widehat{\mathbf{B}}_r^{\mathrm{T}}\boldsymbol{\mu}_x - \hat{\boldsymbol{\mu}}_{r_{y\pi}})(1-p) + N^{-1}\widehat{\mathbf{B}}_r^{\mathrm{T}}\left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}\right) + N^{-1}(\mathbf{Y}_r - \hat{\mathbf{Y}}_r)^{\mathrm{T}}\mathbf{D}_r\boldsymbol{\theta}_r \tag{3.11}$$

Proof: If the calibration equation has a solution $\boldsymbol{\lambda}$, then from (3.5) the difference between the general calibration estimator and the Horvitz-Thompson estimator can be written as:

$$\hat{t}_{r_{yw}} - \hat{t}_{r_{y\pi}} = \sum_r d_k y_k \{q_k \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r + \theta_k(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r)\} \tag{3.12}$$

From (3.7),

$$\boldsymbol{\lambda}_r = \mathbf{T}_r^{-1}\left(\mathbf{t}_x - \mathbf{t}_{r_x}\right) + \mathbf{T}_r^{-1}\left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}\right) - \mathbf{T}_r^{-1}\sum_r d_k \mathbf{x}_k \theta_k\left(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r\right) \tag{3.13}$$

Replacing $\boldsymbol{\lambda}_r$ in (3.12) by the right-hand side of (3.13), we obtain

$$\begin{aligned}
&\hat{t}_{r_{yw}} - \hat{t}_{r_{y\pi}} \\
&= \sum_r d_k y_k \{q_k \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r + \theta_k(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r)\} \\
&= \sum_r d_k y_k \left\{ q_k \mathbf{x}_k^{\mathrm{T}} \left\{ \mathbf{T}_r^{-1}\left(\mathbf{t}_x - \mathbf{t}_{r_x}\right) + \mathbf{T}_r^{-1}\left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}\right) - \mathbf{T}_r^{-1}\sum_r d_k \mathbf{x}_k \theta_k\left(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r\right) \right\} + \theta_k(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r) \right\} \\
&= \sum_r d_k y_k q_k \mathbf{x}_k^{\mathrm{T}} \mathbf{T}_r^{-1}\left(\mathbf{t}_x - \mathbf{t}_{r_x}\right) + \sum_r d_k y_k q_k \mathbf{x}_k^{\mathrm{T}} \mathbf{T}_r^{-1}\left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}\right) - \sum_r d_k y_k q_k \mathbf{x}_k^{\mathrm{T}} \mathbf{T}_r^{-1} \sum_r d_k \theta_k\left(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r\right) \mathbf{x}_k + \sum_r d_k y_k \theta_k(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r) \\
&= \left(\mathbf{X}_r^{\mathrm{T}} \mathbf{D}_r \mathbf{Q}_r \mathbf{Y}_r\right)^{\mathrm{T}} \mathbf{T}_r^{-1}\left(\mathbf{t}_x - \mathbf{t}_{r_x}\right) + \left(\mathbf{X}_r^{\mathrm{T}} \mathbf{D}_r \mathbf{Q}_r \mathbf{Y}_r\right)^{\mathrm{T}} \mathbf{T}_r^{-1}\left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}\right) - \left(\mathbf{X}_r^{\mathrm{T}} \mathbf{D}_r \mathbf{Q}_r \mathbf{Y}_r\right)^{\mathrm{T}} \mathbf{T}_r^{-1} \mathbf{X}_r^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r + \mathbf{Y}_r^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r \\
&= \left(\mathbf{X}_r^{\mathrm{T}} \mathbf{D}_r \mathbf{Q}_r \mathbf{Y}_r\right)^{\mathrm{T}} \mathbf{T}_r^{-1}\left(\mathbf{t}_x - \mathbf{t}_{r_x}\right) + \left(\mathbf{X}_r^{\mathrm{T}} \mathbf{D}_r \mathbf{Q}_r \mathbf{Y}_r\right)^{\mathrm{T}} \mathbf{T}_r^{-1}\left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}\right) - \hat{\mathbf{B}}_r^{\mathrm{T}} \mathbf{X}_r^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r + \mathbf{Y}_r^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r \\
&= \hat{\mathbf{B}}_r^{\mathrm{T}}\left(\mathbf{t}_x - \mathbf{t}_{r_x}\right) + \hat{\mathbf{B}}_r^{\mathrm{T}}\left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}\right) - \hat{\mathbf{Y}}_r^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r + \mathbf{Y}_r^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r \\
&= \hat{\mathbf{B}}_r^{\mathrm{T}}\left(\mathbf{t}_x - \mathbf{t}_{r_x}\right) + \hat{\mathbf{B}}_r^{\mathrm{T}}\left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}\right) + (\mathbf{Y}_r - \hat{\mathbf{Y}}_r)^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r
\end{aligned} \tag{3.14}$$

Then the difference between two means is

$$\begin{aligned}
&N^{-1}\hat{t}_{r_{yw}} - N_r^{-1}\hat{t}_{r_{y\pi}} \\
&= N^{-1}\sum_r d_k y_k \{q_k \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r + \theta_k(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r)\} + (N^{-1} - N_r^{-1})\hat{t}_{r_{y\pi}} \\
&= N^{-1}\hat{\mathbf{B}}_r^{\mathrm{T}}\left(\mathbf{t}_x - \mathbf{t}_{r_x}\right) + N^{-1}\hat{\mathbf{B}}_r^{\mathrm{T}}\left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}\right) + N^{-1}(\mathbf{Y}_r - \hat{\mathbf{Y}}_r)^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r + (\hat{t}_{r_{y\pi}} / N_r)(N_r / N - 1) \\
&= \hat{\mathbf{B}}_r^{\mathrm{T}}\left(\mathbf{t}_x / N - (\mathbf{t}_{r_x} / N_r)(N_r / N)\right) + (\hat{t}_{r_{y\pi}} / N_r)(N_r / N - 1) + N^{-1}\hat{\mathbf{B}}_r^{\mathrm{T}}\left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}\right) + N^{-1}(\mathbf{Y}_r - \hat{\mathbf{Y}}_r)^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r \\
&= \hat{\mathbf{B}}_r^{\mathrm{T}}\left(\boldsymbol{\mu}_x - \boldsymbol{\mu}_{r_x}(N_r / N)\right) + \hat{\boldsymbol{\mu}}_{r_{y\pi}}(N_r / N - 1) + N^{-1}\hat{\mathbf{B}}_r^{\mathrm{T}}\left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}\right) + N^{-1}(\mathbf{Y}_r - \hat{\mathbf{Y}}_r)^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r \\
&= \hat{\mathbf{B}}_r^{\mathrm{T}}\left(\boldsymbol{\mu}_x - \boldsymbol{\mu}_x(\boldsymbol{\mu}_{r_x} / \boldsymbol{\mu}_x)p\right) - \hat{\boldsymbol{\mu}}_{r_{y\pi}}(1 - p) + N^{-1}\hat{\mathbf{B}}_r^{\mathrm{T}}\left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}\right) + N^{-1}(\mathbf{Y}_r - \hat{\mathbf{Y}}_r)^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r \\
&= \hat{\mathbf{B}}_r^{\mathrm{T}} \boldsymbol{\mu}_x \left(1 - (\boldsymbol{\mu}_{r_x} / \boldsymbol{\mu}_x)p\right) - \hat{\boldsymbol{\mu}}_{r_{y\pi}}(1 - p) + N^{-1}\hat{\mathbf{B}}_r^{\mathrm{T}}\left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}\right) + N^{-1}(\mathbf{Y}_r - \hat{\mathbf{Y}}_r)^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r
\end{aligned} \tag{3.15}$$

For the right-hand side of (3.15), the first two terms do not cancel out except in some special situations such as $\boldsymbol{\mu}_{r_x} = \boldsymbol{\mu}_x$ (indicating ignorable nonresponse) and $\hat{\mathbf{B}}_r^{\mathrm{T}} \boldsymbol{\mu}_x = \hat{\boldsymbol{\mu}}_{r_{y\pi}}$ (meaning that the assisting linear regression model has perfect predicting power). The third term is $O_p\left(n_r^{-1/2}\right)$. Based on the New Result 3, we know $\boldsymbol{\theta}_r = O_p(1)$, so the fourth term does not necessarily diminish as n_r increases. Instead, its magnitude seems to depend on the variation of the outcome variable, the predicting power of the regression model underlying the GREG estimator, and the form of the weight adjustment function used in calibration. In general, the difference between general calibration estimator and Horvitz-Thompson estimator does not decrease as n_r increases.

New Result 5. The difference between the general calibration estimator and the GREG estimator is $N^{-1}\left(\hat{t}_{r_{yw}}-\hat{t}_{r_{yreg}}\right)=N^{-1}\mathbf{Y}_r^{\mathrm{T}}\mathbf{D}_r\boldsymbol{\theta}_r=O_p(1)$.

Proof: From (3.5) and (3.14), the general calibration estimator can be expressed as:

$$\begin{aligned}
\hat{t}_{r_{yw}} &= \sum_r d_k y_k + \sum_r d_k q_k \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r y_k + \sum_r d_k\theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r\right)y_k \\
&= \sum_r d_k y_k + \sum_r d_k q_k \mathbf{x}_k^{\mathrm{T}}\left\{\mathbf{T}_r^{-1}\left(\mathbf{t}_x-\mathbf{t}_{r_x}\right)+\mathbf{T}_r^{-1}\left(\mathbf{t}_{r_x}-\hat{\mathbf{t}}_{r_x}\right)-\mathbf{T}_r^{-1}\sum_r d_k\mathbf{x}_k\theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r\right)\right\}y_k+\sum_r d_k\theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r\right)y_k \\
&= \sum_r d_k y_k + \sum_r d_k q_k \mathbf{x}_k^{\mathrm{T}}\mathbf{T}_r^{-1}\left(\mathbf{t}_x-\mathbf{t}_{r_x}\right)y_k+\sum_r d_k q_k \mathbf{x}_k^{\mathrm{T}}\mathbf{T}_r^{-1}\left(\mathbf{t}_{r_x}-\hat{\mathbf{t}}_{r_x}\right)y_k-\sum_r d_k q_k \mathbf{x}_k^{\mathrm{T}}\mathbf{T}_r^{-1}\sum_r d_k\mathbf{x}_k\theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r\right)y_k+\sum_r d_k\theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r\right)y_k \\
&= \hat{t}_{r_{y\pi}}+\widehat{\mathbf{B}}_r^{\mathrm{T}}\left(\mathbf{t}_x-\mathbf{t}_{r_x}\right)+\widehat{\mathbf{B}}_r^{\mathrm{T}}\left(\mathbf{t}_{r_x}-\hat{\mathbf{t}}_{r_x}\right)-\hat{\mathbf{Y}}_r^{\mathrm{T}}\mathbf{D}_r\boldsymbol{\theta}_r+\mathbf{Y}_r^{\mathrm{T}}\mathbf{D}_r\boldsymbol{\theta}_r
\end{aligned} \tag{3.16}$$

But the first four terms of the right-hand side of (3.16) is the GREG estimator

$$\hat{t}_{r_{yreg}}=\hat{t}_{r_{y\pi}}+\widehat{\mathbf{B}}_r^{\mathrm{T}}\left(\mathbf{t}_x-\mathbf{t}_{r_x}\right)+\widehat{\mathbf{B}}_r^{\mathrm{T}}\left(\mathbf{t}_{r_x}-\hat{\mathbf{t}}_{r_x}\right)-\hat{\mathbf{Y}}_r^{\mathrm{T}}\mathbf{D}_r\boldsymbol{\theta}_r \tag{3.17}$$

So $N^{-1}\left(\hat{t}_{r_{yw}}-\hat{t}_{r_{yreg}}\right)=N^{-1}\mathbf{Y}_r^{\mathrm{T}}\mathbf{D}_r\boldsymbol{\theta}_r=O_p(1)$.

The term $\boldsymbol{\theta}$ captures the difference between the weight adjustment function for any general calibration estimator and the weight adjustment function for the GREG estimator. When calibration is used for nonresponse adjustment, $\boldsymbol{\theta}=O_p(1)$ in general situations and does not tend to zero as the sample size n_r increases. As a result, the GREG estimator and the general calibration estimator are not asymptotically equivalent.

The results in this section are purely design-based and provide some initial insight on the difference between the general calibration estimator and GREG estimator when calibration is used for nonresponse adjustment. To gauge the magnitude of the difference, we need to go beyond the design-based approach and examine the underlying models for the population structure and response mechanism. For example, a set of variables may be correlated with the outcome variable of interest. Another set of variables may be correlated with response propensity. The question is how to incorporate these covariates in the calibration process to reduce potential nonresponse bias without increasing variance significantly.

4. Comparison of Three Commonly Used Calibration Estimators for Nonresponse Adjustment through Simulation Study

In this section, we focus on three commonly used calibration estimators in the situation where the auxiliary information is in the form of counts in a frequency table in two or more dimensions. We examine raking (as an example of the general calibration estimators), poststratification (as a special form of the GREG estimator that accounts for the interaction effects of the auxiliary variables), and the GREG estimator that accounts for only the main effects of the auxiliary variables. In practice, the choice between these estimators is often based on the distribution of respondents in the analytic survey and the availability of external data. This section develops a systematic approach for evaluating the performance of these estimators through a simulation study. We compare the unconditional and conditional empirical biases, empirical variances, and the coverage rates of 95 percent confidence intervals of these estimators. The findings demonstrate the importance of accounting for the outcome variable model and response model when choosing the appropriate calibration estimator. A framework involving both design-based and model-based thinking is developed to simultaneously evaluate the impact of sampling, outcome variable structure, and nonresponse mechanism.

Since survey practitioners often lack the knowledge of the outcome variables and nonresponse mechanism in a real-world survey, we also develop a diagnostic method that helps gauge the potential consequence of failure to incorporate significant covariates in the calibration process. The results of this section provide survey practitioners with guidelines for choosing between these commonly used calibration estimators.

4.1 Outcome Variable Model, Response Model, and Covariates in Weighting

The results in Section 3 indicate that the GREG estimator and general calibration estimators are not necessarily asymptotically equivalent when calibration is used for nonresponse adjustment. To gauge how much a general calibration estimator may diverge from a GREG estimator in a particular setting, it is necessary to go beyond the design-based approach and examine the underlying models for the outcome variable and response mechanism. For example, a set of covariates $\mathbf{X}_1$ may determine the outcome variable of interest, while a set of covariates $\mathbf{X}_2$ may drive the response propensity. The relationship between $\mathbf{X}_1$ and $\mathbf{X}_2$ can fall into one of the three situations: i) $\mathbf{X}_1$ and $\mathbf{X}_2$ are exactly the same; ii) $\mathbf{X}_1$ and $\mathbf{X}_2$ are different but have overlapping components; and iii) $\mathbf{X}_1$ and $\mathbf{X}_2$ are completely different with no overlapping components. In practice, we often face the situations that some the covariates corresponding to the response propensity are not correlated with the outcome (situation ii), so the question is whether and how to incorporate these covariates in the calibration process to reduce nonresponse bias without increasing variance significantly.

Based on Little and Vartivarian (2006), "the most important feature of variables for inclusion in weighing adjustments is that they are predictive of survey outcome; prediction of propensity to respond is secondary, although useful." As shown in Table 1 (which is reproduced from Table 1 in their paper), Little and Vartivarian (2006) assess four scenarios based on the association of the auxiliary variables with response and outcome, and draw the following conclusions:

L&V (i): Substantial bias reduction requires adjustment cell variables that are related both to nonresponse and to the outcome of interest.

L&V (ii): If the adjustment cell variables are unrelated to nonresponse, then weighting tends to have no impact on bias, but reduces variance to the extent that the adjustment cell variables are good predictors of the outcome.

L&V (iii): If adjustment cell variables are good predictors of nonresponse but unrelated to the outcome variable, then weighting increases variance without any reduction in bias.

L&V (iv): If the adjustment cell variables are related to neither outcome nor nonresponse, then weighting affects neither bias nor variance.

Table 1. Summary of Little and Vartivarian (2006) Conclusions

Scenario	Association with Outcome	Association with Response	Bias	Variance
L&V (i)	High	High	↓	↓
L&V (ii)	High	Low	--	↓
L&V (iii)	Low	High	--	↑
L&V (iv)	Low	Low	--	--

SOURCE: Little and Vartivarian (2006), Table 1.

In the single-step weighting approach, calibration is applied to the basic sampling weights directly without a separate nonresponse adjustment step, so Little and Vartivarian (2006) offer a useful framework for choosing auxiliary variables and corresponding calibration estimator (e.g., poststratification versus raking). However, the messages in Little and Vartivarian (2006) are not quite clear to the reader sometimes. For example, on the one hand, they assert that "[a] covariate for a weighting adjustment must have two characteristics to reduce nonresponse bias – it needs to be related to the probability of response, and it needs to be related to the survey outcome." On the other hand, they state that "the most important feature of variables for inclusion in weighting adjustment is that they are predictive of survey outcome; prediction of propensity to respond is a secondary, though useful, goal." The former statement seems to suggest that the outcome variable model and response model should play equally important roles in determining the appropriate covariates for nonresponse adjustment, while the latter seems to indicate that the outcome variable model should be the dominant factor. Moreover, the wordings in their text and their Table 1 are not quite consistent. The text seems to address extreme conditions where the variables are either "related" or "not related" to the outcome and/or response, while Table 1 shows "high" and "low" correlations, which are the middle-ground conditions that we are more likely to see in reality. Finally, Little and Vartivarian (2006) address only main

effects and do not provide any guidance about how to handle the interaction effects. Since the interaction terms of the main effect variables are not completely new variables, the conclusions in Little and Vartivarian (2006) cannot be applied directly to the comparison between poststratification, raking, and the GREG with only the main effect terms. In our research, we attempt to address the issues related to the interaction terms and refine the conclusions in Little & Vartivarian (2006) through a simulation study.

4.2 Poststratification, Raking, and the GREG without Interaction Effects

In the simulation study, we focus on three widely used calibration estimators: (1) poststratification estimator as a special case of the GREG estimator, where both main and interaction effects of the categorical auxiliary variables are taken account of; (2) raking ratio estimator as an example of the general calibration estimators; and (3) the GREG estimator when only the main effects of the auxiliary variables are accounted for. For simplicity, we refer to the GREG estimator accounting for only the main effects as "GREG_Main".

The poststratification estimator is generated under the group-mean assisting model. The auxiliary information consists of known cell counts in a frequency table in any number of dimensions. For simplicity, we consider a two-way table with r rows and c columns, and thus $r \times c$ mutually exclusive cells. The auxiliary vector $\mathbf{x}_k$ is composed of $rc-1$ entries of 0 and a single entry of 1 indicating the cell to which k belongs. The population cell U_{ij} contains N_{ij} elements, i=1, …, r; j=1, …, c. So $N = \sum_{i=1}^{r}\sum_{j=1}^{c} N_{ij}$. Sometimes we do not have all the cell counts N_{ij}, but only marginal counts for the benchmark controls. One way to utilize the auxiliary information is to calibrate on known marginal counts, referred to as generalized raking (Deville and Särndal, 1993). Deming and Stephan (1943) suggests an iterative proportional fitting procedure that adjusts one marginal at a time until convergence is achieved. An alternative approach to take advantage of the marginal counts is to use the GREG estimator by accounting for only the main effects of the auxiliary variables (i.e., GREG_Main). We conduct pairwise comparison of these three estimators through simulation.

Raking versus GREG_Main: These two calibration estimators share the same set of auxiliary variables, and their difference lies in the form of distance function $G(\cdot)$ and corresponding adjustment function $F(\cdot)$. In the pure sampling context as discussed in Deville and Särndal (1992), these two estimators are asymptotically equivalent. That is, conditioning on the same set of auxiliary variables, the particular form of the distance function has negligible impact on the asymptotic property of the calibration estimator if non-sampling error does not exist. However, the conclusion in Deville and Särndal (1992) does not always hold when nonresponse exists and calibration is used to reduce nonresponse bias. The theoretical results in Section 3 suggest that the difference between raking and GREG_Main could be as large as $O_p(1)$. The question is in what situation the two estimators tend to give very similar results and in what situations they tend to diverge significantly.

GREG_Main versus poststratification: These two estimators both belong to the GREG estimator family, although with poststratification there is a unique solution to the calibration equations regardless of the distance function involved. GREG_Main accounts for only the main effects of the auxiliary variables while poststratification accounts for the interaction effects as well. This comparison demonstrates the impact of outcome variable model and response model in the choice of calibration estimator. We relate our simulation results to Little and Vartivarian (2006) and provide more refined guidelines for choosing auxiliary variables in nonresponse adjustment weighting.

Poststratification versus raking: These are probably the two most commonly used calibration estimators in US government surveys. From the practical perspective, the key difference between poststratification and raking seems to be that the former fits a fully saturated model with both main and interaction effects of the auxiliary variables, while the latter fits a model including only the main effects. On the other hand, Deville, Särndal, and Sautory (1993) refer to poststratification as *complete poststratification* and raking ratio as *incomplete poststratification*. Does this imply that the raking estimator accounts for the interaction effects to some extent? We attempt to investigate to what extent raking can get closer to poststratification compared to what GREG_Main does.

4.3 Scope and Conceptual Framework for Simulation Study

The simulation study aims to evaluate the empirical properties of the poststratification estimator, the raking ratio estimator, and the GREG_Main estimator for finite population totals and means when calibration is used for nonresponse adjustment in a one-step weighting approach. We measure the magnitude of their differences in terms of empirical bias, variance, mean square error (MSE), and coverage rate of the 95 percent confidence interval, under different model assumptions for the outcome variable and nonresponse mechanism. The research is conducted in the following scope.

First, we evaluate estimates for population totals and means for a single outcome variable. In the presence of nonresponse, calibration is used to reduce the bias, variance, and mean squared error (MSE) of the estimate for this single outcome variable.

Second, although Section 2 points out that it is possible to use a covariate vector $\mathbf{z}_k$ for the calibration adjustment function $F(\cdot)$ that is different from the auxiliary vector $\mathbf{x}_k$ in the calibration equation, our evaluation focuses on the situation where $\mathbf{z}_k = \mathbf{x}_k$, which is the "standard choice" in Särndal and Lundström (2005).

Third, the outcome variable model and response model contain the same main effect covariates. We also assume that there are only two main effect covariates and they are both categorical variables. Sometimes a substantively and statistically significant interaction term between the two main effect variables is also included in either or both models. Depending on the particular calibration estimator, the interaction term may or may not be included in the calibration equation.

Fourth, for the response mechanism, we assume missing at random (MAR). This means that the probability of response does not depend on the outcome variable once we control for the known covariates. The classes or cells defined by the covariates are response homogeneity groups.

Finally, the results focus on overall estimates in the context of simple random sampling (SRS). Although practical surveys almost always involve complex sample designs, the SRS assumption allows us to focus on the impact of population structure and response mechanism on the performance of a calibration estimator. The findings about how to choose auxiliary variables and calibration estimators apply in general to complex designs, although the technical details become more complicated.

In the simulation study, the models for the outcome variable and response propensity are varied by either including or excluding the interaction term of the main effect covariates. We then compare the properties of the poststratification estimator, raking ratio estimator, and GREG_Main estimator in one-step calibration weighting. The alternative models for the outcome variable Y and response propensity R are specified as below. Depending on whether the interaction term is included, we refer to the models as "Y_Main" and "R_Main" (meaning only main effects are in the model) versus "Y_Interaction" and "R_Interaction" (meaning that the interaction effects are included in the model as well).

$$\text{Y\_Main:} \quad Y_{ijk} = \mu_Y + \alpha_{Yi} + \beta_{Yj} + \varepsilon_{Yijk}, \qquad i = 1,\ 2; j = 1,\ 2; k = 1,\ \ldots, N_{ij} \tag{4.1}$$

$$\text{Y\_Interaction:} \quad Y_{ijk} = \mu_Y + \alpha_{Yi} + \beta_{Yj} + \gamma_{Yij} + \varepsilon_{Yijk}, \qquad i = 1,\ 2; j = 1,\ 2; k = 1,\ \ldots, N_{ij} \tag{4.2}$$

$$\text{R\_Main:} \quad R_{ijk} = \mu_R + \alpha_{Ri} + \beta_{Rj} + \varepsilon_{Rijk}, \qquad i = 1,\ 2; j = 1,\ 2; k = 1,\ \ldots, N_{ij} \tag{4.3}$$

$$\text{R\_Interaction:} \quad R_{ijk} = \mu_R + \alpha_{Ri} + \beta_{Rj} + \gamma_{Rij} + \varepsilon_{Rijk}, \qquad i = 1,\ 2; j = 1,\ 2; k = 1,\ \ldots, N_{ij} \tag{4.4}$$

where N_{ij} is the population size in cell ij for the survey, $\varepsilon_{Yijk} \sim N(0, \sigma_Y^2)$, and $\varepsilon_{Rijk} \sim N(0, \sigma_R^2)$.

We are aware that in theory, it would be more appropriate to use logistic models for the response propensity, but choose to use linear models in order to manipulate the simulation parameters easily.

Table 2 summarizes the four simulation scenarios and the corresponding models governing the outcome variable and response propensity. E_M means expectation in terms of the underlying superpopulation model that generates the finite population, and E_R means expectation in terms of the response model. A hypothetical example can be that we are interested in a single outcome variable income. Both the outcome variable y and the response indicator r can be explained by two dichotomous variables, education (high and low) and age (young versus old), and possibly an interaction effect term. In our models, $\boldsymbol{\alpha}$ measures the main effect of education, $\boldsymbol{\beta}$ measures the main effect of age, and $\boldsymbol{\gamma}$ measures the possible interaction effect between education and age. In poststratification, the weights are adjusted by four cells defined by education and age. In raking, the weighting adjustment is conducted iteratively by using age and education as marginal controls until convergence is achieved. In GREG_Main, the calibration estimator is a function of the regression coefficient as the result of modeling the outcome variable income by only the main effects of education and age.

Table 2. Simulation Scenarios

Scenario	Outcome Variable Model	Response Propensity Model
Y_Main & R_Main: neither outcome model nor response model includes interaction effect term	$E_M(Y_{ijk}) = \mu_Y + \alpha_{Yi} + \beta_{Yj}$	$E_R(r_{ijk}) = \mu_R + \alpha_{Ri} + \beta_{Rj}$
Y_Main & R_Interaction: only response model includes interaction term	$E_M(Y_{ijk}) = \mu_Y + \alpha_{Yi} + \beta_{Yj}$	$E_R(r_{ijk}) = \mu_R + \alpha_{Ri} + \beta_{Rj} + \gamma_{Rij}$
Y_Interaction & R_Main: only outcome model includes interaction effect term	$E_M(Y_{ijk}) = \mu_Y + \alpha_{Yi} + \beta_{Yj} + \gamma_{Yij}$	$E_R(r_{ijk}) = \mu_R + \alpha_{Ri} + \beta_{Rj}$
Y_Interaction & R_Interaction: both outcome model and response model include interaction effect term	$E_M(Y_{ijk}) = \mu_Y + \alpha_{Yi} + \beta_{Yj} + \gamma_{Yij}$	$E_R(r_{ijk}) = \mu_R + \alpha_{Ri} + \beta_{Rj} + \gamma_{Rij}$

Through the simulation study, we examine the performance of poststratification, raking, and GREG_Main under different outcome variable model and response model combinations. We evaluate the consistency between our results and those in Little & Vartivarian (2006) and refine their conclusions. At the same time, we attempt to expand Deville and Särndal (1992) and shed light on the empirical difference between the GREG estimators (i.e., GREG_Main and poststratification) and the raking estimator (as an example of the general calibration estimator) in the presence of nonresponse.

4.4 Simulation Steps and Parameters

The implementation of the simulation is based on the framework shown in Table 2 and involves the following steps:

1. Generate an artificial finite population of size N that contains 4 subpopulations defined by the categories of the two auxiliary variables (2x2). The subpopulation size, N_{ij}, is generated using Poisson distribution and approximately equal across the four subpopulations.

2. Generate two sets of values for the outcome variable, one set corresponding to the outcome model Y_Main specified in (4.1) and the other set corresponding to the outcome model Y_Interaction specified in (4.2).

3. Select a simple random sample of size n from the finite population.

4. From the SRS sample, draw subsamples of respondents using the response models R_Main and R_Interaction as specified in (4.3) and (4.4). This is achieved through Poisson sampling and the final sample size (of respondents) is n_r.

5. For each of the four respondent samples corresponding to the different outcome variable and response model combinations shown in Table 2, conduct calibration using poststratification, raking, and GREG_Main,

respectively. Obtain the estimates for the outcome variable associated with the three calibration estimators, and then compare the empirical results using the evaluation criteria specified in Section 4.5.

Several factors may affect the properties of and differences between these three calibration estimators, including: (1) the number of simulation samples; (2) the substantive and statistical significance of the interaction effect in the outcome variable model; (3) the substantive and statistical significance of the interaction effect in the response model; and (4) the overall sample size for the respondent sample and the distribution across the four subpopulations. During the initial investigation, we choose the simulation parameters in a way to minimize the factors that could cloud our comparison.

First, the number of simulation samples is very large ($s = 1,\ 2,\ ...,\ 10000$). That is, for each outcome variable model and response model combination, we draw 10,000 SRS samples from the finite population and then 10,000 corresponding respondent subsamples. The large number of repetitions ensures that any observed differences in the three calibration estimators are not due to the random errors in the simulation samples. This is particularly important when we partition the samples into groups and evaluate the conditional properties of the calibration estimators.

Second, for all the outcome variable and response propensity models (regardless of whether the interaction effect is included or not), the random error terms are set to be very small, so the explanatory power of the overall model is strong. At the same time, the interaction terms in the Y_Interaction model and R_Interaction model are set to be substantively and statistically significant. Such setup ensures that any significant impact of the outcome variable model and response model on the properties of the calibration estimators can be detected. Under these criteria, several sets of outcome model parameters and response model parameters are used. The results associated with these different parameters tend to lead to the same conclusions, so we choose to present the results based on only one set of parameters, as shown below.

The parameters corresponding to the outcome variable models (4.1) and (4.2) are

$$\mu_Y = 1000$$
$$\boldsymbol{\alpha}_Y = (\alpha_{Y1},\ \alpha_{Y2}) = (-200,\ 300)$$
$$\boldsymbol{\beta}_Y = (\beta_{Y1},\ \beta_{Y2}) = (-100,\ 150)$$
$$\boldsymbol{\gamma}_Y = (\gamma_{Y11,}\ \gamma_{Y12,}\gamma_{Y21,}\gamma_{Y22}) = (100,\ 300,\ 700,\ 1200)$$
$$\varepsilon_{Yijk} \sim N(0, 900)$$

The parameters corresponding to the response propensity models (4.3) and (4.4) are

$$\mu_R = 0.05$$
$$\boldsymbol{\alpha}_R = (\alpha_{R1},\ \alpha_{R2}) = (0.05,\ 0.1)$$
$$\boldsymbol{\beta}_R = (\beta_{R1},\ \beta_{R2}) = (0.2,\ 0.4)$$
$$\boldsymbol{\gamma}_R = (\gamma_{R11,}\ \gamma_{R12,}\gamma_{R21,}\gamma_{R22}) = (0.05,\ 0.15,\ 0.2,\ 0.4)$$
$$\varepsilon_{Rijk} \sim N(0, 0.0025)$$

Tables 3 and 4 show the expected cell means corresponding to the models with and without interaction terms, for the outcome variable and the response propensity respectively.

Table 3. Expected Cell Means for Outcome Variable

Outcome Variable Model	$E_M(y_{11k})$	$E_M(y_{12k})$	$E_M(y_{21k})$	$E_M(y_{22k})$
Y_Main	700	950	1,200	1,450
Y_Interaction	800	1,250	1,900	2,650

Table 4. Expected Cell Means for Response Propensity

Response Propensity Model	$E_R(r_{11k})$	$E_R(r_{12k})$	$E_R(r_{21k})$	$E_R(r_{22k})$
R_Main	0.30	0.50	0.35	0.55
R_Interaction	0.35	0.65	0.55	0.95

Finally, the respondent sample size is determined by the SRS sample size n as well as response propensity. Under a particular set of parameters for the response propensity model, we vary the respondent sample size by changing the SRS sample size n. Comparing the results corresponding to large ($n = 8,000$), medium ($n = 2,000$), and small ($n = 200$) respondent sample sizes allows us to evaluate asymptotic properties of these three calibration estimators.

The simulation is conducted in R (Lumley, 2005; R Development Core Team, 2005) because of its efficiency in handling matrix calculations and extensive capacity for analyzing survey data.

4.5 Evaluation Criteria

We first examine the empirical properties of the three calibration estimators using repeated sampling approach (i.e., averaging across the 10,000 simulation samples). Then we compare the results conditioning on the types of samples defined by a proposed distance measure. In real-world survey practice, only one sample can be fielded and all the estimates are based on that particular sample. The conditional properties of these calibration estimators help shed light on the importance of choosing the appropriate estimator based on the particular sample a survey practitioner may obtain.

The empirical results for the three calibration estimators are compared under the four outcome variable and response model combinations using several measures across the simulation samples. The measures are described below in terms of totals. We also evaluate the properties of the means using a similar set of measures.

1. Relative bias $RelBias(\hat{t}_{yw_s}) = (1/S)\sum_{s=1}^{S}(\hat{t}_{yw_s} - t_y)\Big/t_y$
   where s indicates a particular sample, S is the total number of samples included. t_y is the true population total, and $\hat{t}_{yw_s}$ is the estimate from sample s using one of the three calibration estimators.

2. Relative standard error $RelSE(\hat{t}_{yw_s}) = \sqrt{var(\hat{t}_{yw_s})}\Big/t_y = \sqrt{(1/S)\sum_{s=1}^{S}\left(\hat{t}_{yw_s} - E_p(\hat{t}_{yw_s})\right)^2}\Big/t_y$
   where $E_p(\hat{t}_{yw_s}) = (1/S)\sum_{s=1}^{S}\hat{t}_{yw_s}$, the expected value of $\hat{t}_{yw_s}$ over repeated sampling.

3. "Relative" square root of MSE
   $RelRMSE(\hat{t}_{yw_s}) = \sqrt{\text{MSE}(\hat{t}_{yw_s})}\Big/t_y = \sqrt{(1/S)\sum_{s=1}^{S}\left(\hat{t}_{yw_s} - E_p(\hat{t}_{yw_s})\right)^2 + \left((1/S)\sum_{s=1}^{S}(\hat{t}_{yw_s} - t_y)\right)^2}\Big/t_y$
4. Coverage rate of the 95 percent confidence interval
   $(1/S)\sum_{s=1}^{S} I\left(\left|\hat{z}_j\right| \le z_{1-\alpha/2}\right)$, where $\alpha = 0.05$ and $\hat{z}_j = (\hat{t}_{yw_s} - t_y)\Big/\sqrt{var(\hat{t}_{yw_s})}$
   where $\alpha = 0.05$ and $\hat{z}_j = (\hat{t}_{yw_s} - t_y)\Big/\sqrt{var(\hat{t}_{yw_s})}$.

5. Bias ratio: calculated as the ratio of $Bias(\hat{t}_{yw_s})$ and square root of $var(\hat{t}_{yw_s})$:

   $$BiasRatio(\hat{t}_{yw_s}) = (1/S)\sum_{s=1}^{S}(\hat{t}_{yw_s} - t_y)\Big/\sqrt{(1/S)\sum_{s=1}^{S}\left(\hat{t}_{yw_s} - E_p(\hat{t}_{yw_s})\right)^2}$$

4.6 Simulation Results

We evaluate the performance of poststratification, GREG_Main and raking first over repeated sampling, and then conditioning on sample types defined by a distance measure. The results for the total and those for the mean demonstrate almost exactly the same pattern, so we focus on discussing the properties of the estimators for the total.

4.6.1 Properties over Repeated Sampling

The empirical bias and variance used for calculating the measures in Table 5 were estimated by averaging across the 10,000 simulation samples. For each of the two outcome variable models (Y_Main and Y_Interaction), we first obtain the results corresponding to 100 percent response as the baseline, and then evaluate the properties of the three calibration estimators under the R_Main and R_Interaction models.

Impact of Outcome Variable Model and Response Propensity Model

Table 5 shows that the relative biases associated with all the three calibration estimators are very small, indicating that calibration helps reduce nonresponse bias significantly so long as the main effects of the key auxiliary variables are accounted for. At the same time, the outcome variable model seems to be the driving factor that determines the performance of the calibration estimators, not the response propensity model. When the outcome variable model does not contain the interaction effect of the auxiliary variables (in the Y_Main scenarios), poststratification (which accounts for the interaction effect) does not seem to reduce nonresponse bias or variance further than raking or GREG_Main, and this is true regardless of the response model (R_Main or R_Interaction). For example, in the "$n = 8,000$, Y_Main & R_Interaction" scenario, the three calibration estimators have almost the same relative bias (approximately 1.8E-5), relative standard error (approximately 4.0E-4), relative square root of MSE (approximately 0.00040), and coverage rate of 95% confidence interval (93 percent).

On the other hand, when the outcome variable model contains the interaction effect term, poststratification performs better in terms of both relative bias and relative standard error regardless of the response propensity model. In the "$n = 8,000$, Y_Interaction & R_Interaction" scenario, the relative biases for raking and GREG_Main are 65 times and 300 times as large as that for poststratification, and the relative standard errors for raking and GREG_Main are approximately 3 times as large as that for poststratification.

We think the key to understanding this pattern of results is the following:If an auxiliary variable is correlated only to nonresponse but uncorrelated to the outcome variable, the auxiliary variable does not cause any nonresponse bias. In the Y_main scenarios, although the interaction term affects nonresponse, such nonresponse does not introduce any nonresponse bias in addition to the nonresponse bias that is already caused by the main effects (because the interaction effect itself is not correlated with the outcome variable). Including the interaction term in calibration does not help because no bias is caused by the interaction term in the first place. This is why raking and GREG_Main performed almost as well as poststratification in the "Y_main" scenarios. The interaction effect in the response model does seem to play a role in the performance of the three calibration estimators, conditioning on the fact that the interaction effect is corrected to the outcome variable. Although GREG_Main performs poorly in the "Y_Interaction & R_Main" scenarios, its worst performance is observed in the "Y_Interaction & R_Interaction" scenarios when the interaction term is related to both the outcome variable and the response propensity.

Coverage Rate of 95% Confidence Interval and Bias Ratio

Despite the differences among the three calibration estimators in the Y_Interaction scenarios, the relative biases are very small even for the calibration estimator that fails to account for the interaction effect that is correlated to the outcome variable. However, the small relative bias can be misleading because the coverage rate of 95% confidence interval can be poor for GREG_Main and raking even with very small relative bias. A very interesting pattern in Table 5 is that as the SRS sample size increases, the confidence interval coverage rates actually become worse for raking and GREG_Main. For example, in the "Y_Interaction & R_Interaction" scenario, the coverage rate of 95% confidence interval for GREG_Main is only 55% with the sample size of 2,000, and it drops further to 3% when the sample size increases to 8,000!

How can it be possible that the coverage rate of 95% confidence interval becomes almost unacceptable even when the relative bias is very low? Why does the increase in sample size hurt the confidence interval coverage rate? The answer lies in the asymptotic property of the bias ratio. We can re-write the *t*-statistic into the summation of two terms:

Table 5. Properties of Poststratification, GREG_Main, and Raking over Repeated Sampling

	Relative Bias $RelBias(\hat{t}_{yw_s})$ in 10^{-5}				Relative Standard Error $RelSE(\hat{t}_{yw_s})$ in 10^{-4}				Relative Square Root of MSE $RelRMSE(\hat{t}_{yw_s})$ in 10^{-4}			Coverage of 95% Confidence Interval		
	No calibration	Poststratification	GREG_Main	Raking	No calibration	Poststratification	GREG_Main	Raking	Poststratification	GREG_Main	Raking	Poststratification	GREG_Main	Raking
SRS sample $n = 8,000$														
Y_Main &100% response	-0.9	-0.1	-0.1	-0.1	26.1	2.8	2.8	2.8	2.2	2.2	2.2	95%	95%	95%
Y_Main & R_Main	4100.0	1.1	1.2	1.2	42.0	4.8	4.8	4.8	3.8	3.8	3.8	93%	93%	93%
Y_Main & R_Interaction	7390.0	1.8	1.9	1.8	33.6	4.0	4.0	4.0	3.2	3.2	3.2	93%	93%	93%
Y_Interaction & 100% response	0.8	-0.1	-0.3	-0.3	39.4	1.8	4.9	4.9	1.5	3.9	3.9	95%	95%	95%
Y_Interaction & R_Main	8660.0	1.1	-135.0	-67.9	65.6	3.1	9.0	8.3	2.5	14.0	8.7	94%	63%	88%
Y_Interaction & R_Interaction	13900.0	1.0	-304.0	-64.5	52.8	2.6	8.2	7.0	2.1	30.4	7.8	94%	3%	90%
SRS sample $n = 2,000$														
Y_Main &100% response	-0.8	-0.7	-0.7	-0.7	56.5	6.1	6.1	6.1	4.9	4.9	4.9	95%	95%	95%
Y_Main & R_Main	4100.0	0.4	0.5	0.4	88.3	9.8	9.8	9.8	7.8	7.8	7.8	95%	95%	95%
Y_Main & R_Interaction	7370.0	2.1	2.3	2.2	70.0	8.5	8.4	8.5	6.7	6.7	6.7	94%	94%	94%
Y_Interaction & 100% response	1.5	-0.2	-0.6	-0.6	87.0	3.9	10.8	10.8	3.2	8.6	8.6	95%	95%	95%
Y_Interaction & R_Main	8670.0	2.2	-136.0	-68.0	136.0	6.4	18.6	17.3	5.1	18.6	14.8	95%	88%	95%
Y_Interaction & R_Interaction	13900.0	1.4	-304.0	-63.9	110.0	5.4	17.0	14.7	4.3	30.9	12.8	95%	55%	96%
SRS sample $n = 200$														
Y_Main &100% response	-30.2	2.8	2.7	2.7	185.0	20.0	20.0	20.0	15.9	15.9	15.9	94%	94%	94%
Y_Main & R_Main	4130.0	-2.1	-1.9	-2.1	280.0	31.9	31.7	31.7	25.5	25.3	25.3	94%	94%	94%
Y_Main & R_Interaction	7400.0	3.1	3.7	3.2	230.0	27.2	26.9	27.0	21.6	21.4	21.4	94%	94%	94%
Y_Interaction & 100% response	3.6	0.0	1.9	2.0	284.0	12.9	35.2	34.8	10.3	28.1	27.8	95%	95%	95%
Y_Interaction & R_Main	8710.0	3.4	-149.0	-77.0	433.0	21.0	62.2	56.4	16.8	50.7	45.2	94%	94%	96%
Y_Interaction & R_Interaction	13900.0	-3.5	-314.0	-72.7	356.0	17.8	56.3	48.4	14.2	50.8	38.7	94%	92%	98%

$$t\text{-}statistic = \frac{\hat{t}_{yw_s} - t_y}{\sqrt{\text{var}(\hat{t}_{yw_s})}} = \frac{\hat{t}_{yw_s} - \text{E}_\text{p}\text{E}_\text{M}(\hat{t}_{yw_s})}{\sqrt{\text{var}(\hat{t}_{yw_s})}} + \frac{\text{E}_\text{p}\text{E}_\text{M}(\hat{t}_{yw_s}) - t_y}{\sqrt{\text{var}(\hat{t}_{yw_s})}} \quad (4.5)$$

The first term on the right-hand side of (4.5) is asymptotically $N(0, 1)$. The second term is the standardized bias or bias ratio. As the sample size increases, the denominator of the second term decreases. However, if the calibration estimator is model-biased as in the situation of GREG_Main and raking, the numerator in the second term of does not decrease with increase sample size, but rather stays constant. As a result, the increase in sample size leads to an increase in the bias ratio and therefore makes the coverage rate of 95% confidence interval worse. An important lesson here is that increasing sample sizes does not help improve the performance of a calibration estimator that is model-biased.

Why Does Raking Perform Better Than GREG_Main?

Another interesting pattern in Table 5 is that in the "Y_Interaction" scenarios, the raking estimator has much smaller relative bias and better confidence interval coverage than GREG_Main. Neither approach takes account of interaction effects during the calibration process, so the question is why raking performs better than GREG_Main.

An important feature of raking is that the algorithm forces the weights to conform to the marginal totals without perturbing the associations in the unadjusted table (Haberman, 1979), so the method preserves the interaction effect that already exists in the data before calibration. That is, raking retains the cross-product ratios or odds ratios of the cell totals in the observed data (Brick, Montaquila, and Roth 2003). In our situation, we conduct raking on the respondent sample using iterative proportional fitting, and this procedure does not change the existing association between the two main effect variables. As shown in Table 6, the odds ratios in the raking column are the same as those in the respondent sample column. In contrast, GREG_Main fits a regression model using only the main effect variables, and it destroys the existing correlation between the two main effect variables in the respondent sample.

Raking and GREG_Main both account for only main effects, but are associated with different distance functions. This comparison shows that the form of distance function matters.

Table 6. Impact of Poststratification, GREG_Main, and Raking on Odds Ratios of Cell Totals

	Odds Ratio of Cell Totals				
SRS sample size = 8,000	Population	Respondent sample	Poststratification	GREG_Main	Raking
R_Main scenarios, regardless of Y model	0.99	0.93	0.99	0.88	0.93
R_Interaction scenarios, regardless of Y model	0.99	0.93	0.99	0.76	0.93

6.2 Properties Conditioning on Sample

The relative biases of raking and GREG_Main seem acceptable over repeated sampling. However, a survey practitioner can obtain one and only one sample in the real world, so it is important to understand how a calibration estimator may perform for a particular sample. We define a measure that helps survey practitioners gauge the potential consequence for choose the inappropriate calibration estimator for a particular sample. The discussions below focus on the comparison between raking and poststratification. A similar comparison can be conducted between GREG_Main and poststratification.

Assume that the model for the outcome variable Y contains an interaction effect term, as specified in equation (4.2). In the SRS setting, the calibration estimator for the total can be expressed as:

$$\hat{t}_{yw} = \sum_{i=1}^{2}\sum_{j=1}^{2} w_{ij} \sum_{k=1}^{n_{ij}} y_{ijk} \quad (4.6)$$

where w_{ij} is the calibrated weight for a unit in cell ij.

The correct calibration method should be poststratification. If raking is used, then the model expectation of the estimator is:

$$
\begin{aligned}
&\mathrm{E_M}(\hat{t}_{raking}) \\
&= \sum_{i=1}^{2}\sum_{j=1}^{2} w_{ij} \sum_{k=1}^{n_{ij}} \mathrm{E_M}(y_{ijk}) \\
&= \sum_{i=1}^{2}\sum_{j=1}^{2} w_{ij} \sum_{k=1}^{n_{ij}} (\mu_Y + \alpha_{Yi} + \beta_{Yj} + \gamma_{Yij}) \\
&= \mu_Y \sum_{i=1}^{2}\sum_{j=1}^{2} w_{ij} n_{ij} + \sum_{i=1}^{2} \alpha_{Yi} \sum_{j=1}^{2} w_{ij} n_{ij} + \sum_{j=1}^{2} \beta_{Yj} \sum_{i=1}^{2} w_{ij} n_{ij} + \sum_{i=1}^{2}\sum_{j=1}^{2} \gamma_{Yij} w_{ij} n_{ij} \\
&= \mu_Y \hat{N} + \sum_{i=1}^{2} \alpha_{Yi} \hat{N}_{i\cdot} + \sum_{j=1}^{2} \beta_{Yj} \hat{N}_{\cdot j} + \sum_{i=1}^{2}\sum_{j=1}^{2} \gamma_{Yij} \hat{N}_{ij}
\end{aligned}
\tag{4.7}
$$

The model bias of the raking estimator $\hat{t}_{raking}$ is:

$$
\begin{aligned}
&\mathrm{E_M}(\hat{t}_{raking} - t) \\
&= \mu_Y (\hat{N} - N) + \sum_{i=1}^{2} \alpha_{Yi} (\hat{N}_{i\cdot} - N_{i\cdot}) + \sum_{j=1}^{2} \beta_{Yj} (\hat{N}_{\cdot j} - N_{\cdot j}) + \sum_{i=1}^{2}\sum_{j=1}^{2} \gamma_{Yij} (\hat{N}_{ij} - N_{ij})
\end{aligned}
\tag{4.8}
$$

The raking process forces the estimated row totals and column totals to be equal to the control totals, so the first three terms of the right-hand side of equation (4.8) should be approximately zero, leaving the fourth to be the driving term.

In practice, we do not know the values for μ_Y, α_{Yi}, β_{Yj}, and γ_{Yij}. However, if we have the cross-classification and corresponding cell totals for the population, we can compute the difference between $\hat{N}_{ij}$ and N_{ij} for each of the four cells defined by the two auxiliary variables. Then we can define a distance measure as the square root of the summation of the differences across all the cells:

$$
\mathrm{D}_{raking} = \sqrt{\sum_{i=1}^{2}\sum_{j=1}^{2} (\hat{N}_{ij} - N_{ij})^2}
\tag{4.9}
$$

This distance measure is computable for each particular sample and can help predict the model bias of the raking estimator for a particular sample when poststratification is supposed to be the appropriate estimator. We use the "$n = 8{,}000$, Y_Interaction & R_Interaction" scenario to demonstrate how this distance measure can be used – First, we compute the distance measure for each of the 10,000 simulation samples. Then we partition the 10,000 samples into 20 groups based on the magnitude of the distance measure, with approximately 500 samples in each group. Finally, we calculate the average relative biases and average coverage rate of 95% percent confidence interval for each of the 20 groups. The results are presented in Table 7.

Table 7 shows that different samples defined by the magnitude of the distance measure behave differently. As the distance measure increases, the relative bias increases and the coverage rate of 95% confidence interval decreases. The coverage rate of 95% confidence interval becomes drops to below 88 percent for a sample with a distance measure above the 80th percentile. If a survey practitioner happens to obtain a sample with distance measure above the 95th percentile, then coverage rate of 95% confidence interval is only 10 percent. So the impact of choosing the wrong estimator can be detrimental for some "unlucky" samples.

In practice, only one sample is fielded for a survey, so we can calculate a single value for the distance measure instead of multiple values associated with many possible samples. Then the question is how to interpret and make

use of this single value. The next step in our research is to refine the form of the distance measure and derive the statistical distribution for it, so survey practitioners can determine the statistical significance of the value associated with a particular sample based on the distribution of the distance measure.

Table 7. Relative Bias and Confidence Interval Coverage Rate for Raking Estimator Conditioning on Sample, in the "$n = 8,000$, Y_Interaction & R_Interaction" Scenario

Range of D_{raking}	Relative Bias in 10^{-5}			Coverage Rate of 95% Confidence Interval		
	Post-stratification	GREG_Main	Raking	Post-stratification	GREG_Main	Raking
0^{th} – 5^{th} percentile: (0.04,29]	1.5	-228.2	1.0	93%	1%	100%
5^{th} – 10^{th} percentile: (29,59]	0.6	-231.7	-1.4	94%	1%	100%
10^{th} – 5^{th} percentile: (59,88]	2.1	-231.6	-1.3	92%	2%	100%
15^{th} – 20^{th} percentile: (88,117]	1.9	-237.7	-6.4	95%	1%	100%
20^{th} – 25^{th} percentile: (117,148]	0.2	-245.9	-14.5	94%	1%	100%
25^{th} – 30^{th} percentile: (148,180]	2.6	-247.7	-16.6	92%	3%	100%
30^{th} – 35^{th} percentile: (180,210]	3.5	-259.6	-23.9	93%	3%	100%
35^{th} – 40^{th} percentile: (210,240]	0.0	-269.9	-35.3	93%	4%	100%
40^{th} – 45^{th} percentile: (240,270]	-0.4	-272.1	-36.8	94%	5%	100%
45^{th} – 50^{th} percentile: (270,301]	1.1	-287.1	-49.7	93%	6%	100%
50^{th} – 55^{th} percentile: (301,332]	1.0	-296.2	-58.3	96%	4%	100%
55^{th} – 60^{th} percentile: (332,366]	0.1	-307.2	-66.6	93%	6%	99%
60^{th} – 65^{th} percentile: (366,401]	1.1	-318.2	-77.4	93%	3%	99%
65^{th} – 70^{th} percentile: (401,439]	0.6	-325.8	-84.7	93%	5%	98%
70^{th} – 75^{th} percentile: (439,480]	0.1	-340.7	-98.2	94%	3%	97%
75^{th} – 80^{th} percentile: (480,528]	-0.3	-354.2	-109.9	95%	2%	93%
80^{th} – 85^{th} percentile: (528,582]	1.1	-367.5	-119.8	95%	2%	88%
85^{th} – 90^{th} percentile: (582,650]	1.4	-385.3	-136.0	95%	1%	70%
90^{th} – 95^{th} percentile: (650,750]	0.7	-410.4	-157.4	92%	0%	42%
95^{th} – 100^{th} percentile:(750,1450]	0.2	-457.9	-197.7	94%	0%	10%

5. Summary

Our research proves that in general, the GREG estimator and general calibration estimators are not asymptotically equivalent in the presence of nonresponse. Through a simulation study, we demonstrate the importance of accounting for the outcome variable model and response model when choosing the appropriate estimator. Furthermore, we point out that the outcome model should be the dominant factor in determining what covariates should be included in the calibration process.

One of the interesting findings is that small relative bias associated with the inappropriate calibration estimator could result in very poor confidence interval coverage. Increasing sample size only makes the situation worse because the bias tends to remain constant while the variance decreases with increasing sample sizes.

In the real-world survey practice where only a single sample is obtained, a distance measure could help gauge the potential consequence of choosing inappropriate estimator. We plan to conduct further research on this and provide more details in our future work.

References

Brick J.M, Montaquila, J., and Roth, S (2003). Identifying problems with raking estimators. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 710-717.

Chang, T., and Kott, P. (2008). Using Calibration Weighting to Adjust for Nonresponse under a Plausible Model. *Biometrika,* 95, 555-571.

Deming, W.E. (1943). *Statistical Adjustment of Data*. New York: John Wiley & Sons, Inc.

Dever, J. (2008). Sampling Weight Calibration with Estimated Control Totals. (Dissertation)

Dever, J., and Valliant, R. (2010). A Comparison of Variance Estimators for Poststratification to Estimated Control Totals. *Survey Methodology*, 36, 45-56.

Deville, J-C., and Särndal, C-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87(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, 379-399.

Estevao, V.M., and Särndal, C.E. (2006). Survey Estimates by Calibration on Complex Auxiliary Information. *International Statistical Review*, 74, 127-147.

Fuller, W. A. (2000). Two-phase Sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 23-30.

Haberman, S. (1979). *Analysis of Qualitative Data*. Academic Press, Inc. New York.

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

Kott, P.S. (1994). A note on Handling Nonresponse in Surveys. *Journal of the American Statistical Association*, 89, 693-696.

Kott, P. S. (2006). Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors. *Survey Methodology,* 32, 133-142.

Kott, P. S., and Chang, T. (2010). Using Calibration Weighting to Adjust for Nonignorable Unit Nonresponse. *Journal of the American Statistical Association*, 105(491), 1265-1275.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R.* John Wiley & Sons: New Jersey.

Lundström, S., and Särndal, C-E. (1999). Calibration as a Standard Method for Treatment of Nonresponse. *Journal of Official Statistics*, 15, 305-327.

R Development Core Team (2005). *R: A Language and Environment for Statistical Computing.* R Development Core Team, Vienna, Austria. URL http://www.R-project.org

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

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

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

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

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, 117-125.

Thibaudeau, Y. and Slud, E. (2009). Simultaneous Calibration and Nonresponse Adjustment. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 2263-2272.