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

Survey Methodology

A design effect measure for calibration weighting in single-stage samples

by Kimberly A. Henry and Richard Valliant

Release date: December 17, 2015

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

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

Depository Services Program

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

Standards of service to the public

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

Note of appreciation

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

Standard table symbols

The following symbols are used in Statistics Canada publications:

. not available for any reference period
.. not available for a specific reference period
... not applicable
0 true zero or a value rounded to zero
0^s value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p} preliminary
^{r} revised
x suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E} use with caution
F too unreliable to be published
* significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2015
Vol. 41, No. 2, pp. 315-331
Statistics Canada, Catalogue No. 12-001-X

A design effect measure for calibration weighting in single-stage samples

Kimberly A. Henry and Richard Valliant[1]

Abstract

We propose a model-assisted extension of weighting design-effect measures. We develop a summary-level statistic for different variables of interest, in single-stage sampling and under calibration weight adjustments. Our proposed design effect measure captures the joint effects of a non-*epsem* sampling design, unequal weights produced using calibration adjustments, and the strength of the association between an analysis variable and the auxiliaries used in calibration. We compare our proposed measure to existing design effect measures in simulations using variables like those collected in establishment surveys and telephone surveys of households.

Key Words: Auxiliary data; Kish weighting design effect; Spencer design effect; Generalized regression estimator.

1 Introduction

A design effect (deff) in its general form measures the relative increase or decrease in the variance of an estimator due to departures from simple random sampling. Kish (1965) presented the deff as a convenient way of gauging the effect of clustering on an estimator of a mean. Park and Lee (2004) review some of the history behind the formulation and use of deff's. Design effects are especially useful in approximating the total sample size needed in a cluster sample. Clustering usually causes some loss of efficiency and the variance from a simple random sample, which is easy to compute, can be multiplied by a deff to approximate the variance that would be obtained from a cluster sample. This can, in turn, be used to determine the total sample size needed in a cluster sample to achieve a desired level of precision. Later work by Rao and Scott (1984) and others found that more complicated versions of deff's were useful to adjust inferential statistics calculated from complex survey data.

A specialized version of the deff was proposed in Kish (1965) that addressed only the effect of using weights that are not all equal. Kish derived the "design effect due to weighting" for a case in which weights vary for reasons other than statistical efficiency. On the other hand, there are sample designs and estimators where having varying weights can be quite efficient. An establishment survey where population variances of analysis variables differ markedly among industries is one example. Calibrating to population counts can also produce different sized weights but is an essential tool in attempting to correct for coverage errors in some surveys, like ones done by telephone. Spencer (2000) proposed a simple model-assisted approach to estimate the impact on variance of using variable weights in a situation where an analysis variable depends on a single covariate.

The Kish and Spencer measures, reviewed in Section 2, do not provide a summary measure of the impact of the gains in precision that may accrue from sampling with varying probabilities and using a calibration estimator like the general regression (GREG) estimator. While the Kish design effects attempt to measure the impact of variable weights, they are informative only under special circumstances, do not

1. Kimberly A. Henry, U.S. Internal Revenue Service, Washington DC. E-mail: kimberly.a.henry@irs.gov; Richard Valliant, Universities of Michigan & Maryland, College Park MD 20854. E-mail: rvalliant@survey.umd.edu.

account for alternative variables of interest, and can incorrectly measure the impact of differential weighting in some circumstances, facts noted in Kish (1992). Survey practitioners should be cautious when using this measure in informative sampling and estimation schemes in which there exists an intentional relationship between the weights and variables of interest. Spencer's approach holds for with-replacement single-stage sampling for a very simple estimator of the total constructed with inverse-probability weights with no further adjustments. There are also few empirical examples comparing these measures in the literature.

Calibration adjustments are often applied to reduce variances and correct for undercoverage and/or nonresponse in surveys (e.g., Särndal and Lundström 2005; Kott 2009). When the calibration covariates are correlated with the coverage/response mechanism, calibration weights can improve the mean squared error (MSE) of an estimator. In many applications, since calibration involves unit-level adjustments, calibration weights can vary more than the base weights or category-based nonresponse or poststratification adjustments (Kalton and Flores-Cervantes 2003; Brick and Montaquila 2009). Thus, an ideal measure of the impact of calibration weights incorporates not only the correlation between the survey variable of interest y and the weights, but also the correlation between y and the calibration covariates $\mathbf{x}$ to avoid "penalizing" weights for the mere sake that they vary.

In Section 3, we introduce a new design effect measure that accounts for the joint effect of a non-epsem sample design and unequal weight adjustments in the larger class of calibration estimators. It is assumed that a probability sample design is used and that there are no missing data problems that would induce a dependence between sample inclusion and the values of the y's. Our summary measure incorporates the survey variable, using a generalized regression variance to reflect multiple calibration covariates. In Section 4, we apply the estimators in a simulation using variables similar to ones collected in establishment surveys and household surveys done by telephone and demonstrate empirically how the proposed estimator outperforms the existing methods in the presence of unequal calibration weights. Section 5 is a conclusion.

2 Existing methods

In this section, we specify notation and summarize the Kish and Spencer measures. The assumptions used to derive each of these are also presented.

2.1 GREG weight adjustments

Case weights resulting from calibration on benchmark auxiliary variables can be defined with a global regression model for the survey variables (see Kott 2009 for a review). Deville and Särndal (1992) proposed the calibration approach that involves minimizing a distance function between the base weights and final weights to obtain an optimal set of survey weights. Specifying alternative calibration distance functions produces alternative estimators. Suppose that a single-stage probability sample of n units is selected with π_i being the selection probability of unit i and $\mathbf{x}_i$ a vector of p auxiliaries associated with unit i. A least squares distance function produces the *general regression estimator* (GREG):

$$\hat{T}_{\text{GREG}} = \hat{T}_{\text{HT}y} + \hat{\mathbf{B}}^T \left(\mathbf{T}_x - \hat{\mathbf{T}}_{\text{HT}x} \right) = \sum_{i \in s} g_i y_i / \pi_i , \qquad (2.1)$$

where $\hat{T}_{\text{HT}y} = \sum_{i\in s} y_i/\pi_i$ is the Horvitz-Thompson (HT 1952) estimator of the population total of y, $\hat{\mathbf{T}}_{\text{HT}x} = \sum_{i\in s} \mathbf{x}_i/\pi_i$ is the vector of HT estimated totals for the auxiliary variables, $\mathbf{T}_x = \sum_{i=1}^{N} \mathbf{x}_i$ is the corresponding vector of known totals, $\hat{\mathbf{B}} = \mathbf{A}_s^{-1}\mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\boldsymbol{\Pi}_s^{-1}\mathbf{y}_s$ is the regression coefficient, with $\mathbf{A}_s = \mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\boldsymbol{\Pi}_s^{-1}\mathbf{X}_s$, $\mathbf{X}_s^T$ is the matrix of $\mathbf{x}_i$ values in the sample, $\mathbf{V}_{ss} = \text{diag}(v_i)$ is the diagonal of the variance matrix specified under the working model $y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \varepsilon_i, \varepsilon_i \sim (0, v_i)$, and $\boldsymbol{\Pi}_s = \text{diag}(\pi_i)$. In the second expression for the GREG estimator in (2.1), $g_i = 1 + \left(\mathbf{T}_x - \hat{\mathbf{T}}_{\text{HT}x}\right)^T \mathbf{A}_s^{-1}\mathbf{x}_i v_i^{-1}$ is the "$g-$weight," such that the case weights are $w_i = g_i/\pi_i$ for each sample unit i.

The GREG estimator for a total is model-unbiased under the associated working model. The GREG is consistent and approximately design-unbiased when the sample size is large (Särndal, Swensson and Wretman 1992). When the model is correct, the GREG estimator achieves efficiency gains. If the model is incorrect, then the efficiency gains will be dampened (or nonexistent) but the GREG estimator is still approximately design-unbiased. Relevant to this work, the variance of the GREG estimator can be used to approximate the variance of any calibration estimator (Deville and Särndal 1992; Deville, Särndal and Sautory 1993) when the sample size is large. This allows us to produce one design effect measure applicable to all estimators in the family of calibration estimators.

2.2 The direct design-effect measures for single-stage samples

For a given non$-$epsem sample π and estimator $\hat{T}$ for the finite population total T, one definition for the *direct design effect* (Kish 1965) is

$$\text{Deff}\left(\hat{T}\right) = \text{Var}_{\pi}\left(\hat{T}\right)/\text{Var}_{\text{srswr}}\left(\hat{T}_{\text{srswr}}\right) \tag{2.2}$$

where $\hat{T}_{\text{srswr}}$ is the estimator of a total based on a simple random sample selected with replacement (srswr). We refer to this as a "direct" population quantity since it uses theoretical variances in the numerator and denominator. The design effect in (2.2) measures the size of the variance of the estimator $\hat{T}$ under the design π, relative to the variance of the estimator of the same total if a srswr of the same size had been used.

In large samples, we can approximate the variance of any calibration estimator $\hat{T}_{\text{cal}}$ using the approximate variance of the GREG (GREG AV, Särndal et al. 1992; Deville et al. 1993), such that the design effect is

$$\text{Deff}\left(\hat{T}_{\text{cal}}\right) \doteq \text{Var}_{\pi}\left(\hat{T}_{\text{GREG}}\right)/\text{Var}_{\text{srswr}}\left(\hat{T}_{\text{srswr}}\right). \tag{2.3}$$

To estimate these design-effects, we use the appropriate corresponding sample-based variance estimates. Estimates of both measures (2.2) and (2.3) can be produced using conventional survey estimation software. Our proposed design effect is a model-assisted approximation to (2.3).

2.3 Kish's "Haphazard-sampling" design-effect measure for unequal weights

Kish (1965, 1990) proposed the "design effect due to weighting" as a measure to quantify the loss of precision due to using unequal and inefficient weights. For $\mathbf{w} = (w_1, \ldots, w_n)^T$, this measure is

$$\begin{aligned} \text{deff}_K(\mathbf{w}) &= 1 + [\text{CV}(\mathbf{w})]^2 \\ &= \frac{n\sum_{i\in s} w_i^2}{\left[\sum_{i\in s} w_i\right]^2}, \end{aligned} \tag{2.4}$$

where $\text{CV}(\mathbf{w}) = \sqrt{n^{-1}\sum_{i\in s}(w_i - \bar{w})^2 / \bar{w}^2}$ is the coefficient of variation of the weights with $\bar{w} = n^{-1}\sum_{i\in s} w_i$. Expression (2.4) is derived from the ratio of the variance of the weighted survey mean under disproportionate stratified sampling to the variance under proportionate stratified sampling when all stratum unit variances are equal (Kish 1992). With equal stratum variances, sampling with a proportional allocation to strata is optimal, which leads to all units having the same weight.

Kish referred to (2.4) as a measure that is appropriate for "haphazard" weighting in which unequal weights are inefficient. Kish (1992) and Park and Lee (2004) give examples of informative sampling where this measure does not apply. Park and Lee (2004) also demonstrate this measure may not apply equally well to estimators of means and totals.

2.4 Spencer's model-assisted measure for PPSWR sampling

Spencer (2000) derives a design-effect measure to more fully account for the effect on variances of weights that are correlated with the survey variable of interest. The sample is assumed to be selected with varying probabilities and with replacement (denoted as PPSWR sampling here). A particular case of this would be $p_i \propto x_i$, where x_i is a measure of size associated with unit i and p_i is the one-draw probability of selecting unit i. Suppose that p_i is correlated with y_i and that a linear model holds for $y_i : y_i = \alpha + \beta p_i + \varepsilon_i$. If the entire finite population were available, then the ordinary least squares estimates of α and β are $A = \bar{Y} - B\bar{P}$ and $B = \sum_{i\in U}(y_i - \bar{Y})(p_i - \bar{P}) / \sum_{i\in U}(p_i - \bar{P})^2$, where $\bar{Y}, \bar{P}$ are the finite population means for y_i and p_i. The finite population variance of the residuals, $e_i = y_i - (A + Bp_i)$, is $\sigma_e^2 = (1 - \rho_{yp}^2) N^{-1}\sum_{i\in U}(y_i - \bar{Y})^2 = (1 - \rho_{yp}^2)\sigma_y^2$, where σ_y^2 is the finite population variance of y and ρ_{yp} is the finite population correlation between y_i and p_i. The estimated total studied by Spencer is referred to as the pwr – estimator or Hansen-Hurwitz (1943) estimator (Särndal et al. 1992, Section 2.9) and is defined as $\hat{T}_{\text{pwr}} = n^{-1}\sum_{i=1}^{n} y_i / p_i$, with design-variance $\text{Var}(\hat{T}_{\text{pwr}}) = n^{-1}\sum_{i\in U} p_i (y_i / p_i - T)^2$ in single-stage sampling. For use below, define $w_i = (np_i)^{-1}$. Spencer substituted the model-based values for y_i into the pwr – estimator's variance and took its ratio to the variance of the estimated total using srswr to produce the following design effect for unequal weighting (see Appendix in Spencer 2000):

$$\text{Deff}_S = \frac{A^2}{\sigma_y^2}\left(\frac{n\bar{W}}{N} - 1\right) + \frac{n\bar{W}}{N}(1 - \rho_{yp}^2) + \frac{n\rho_{e^2w}\sigma_{e^2}\sigma_w}{N\sigma_y^2} + \frac{2An\rho_{ew}\sigma_e\sigma_w}{N\sigma_y^2} \tag{2.5}$$

where $\bar{W} = N^{-1}\sum_{i\in U} w_i = (nN)^{-1}\sum_{i\in U} 1/p_i$ is the average weight in the population, ρ_{e^2w} and ρ_{ew} are the finite population correlation of the e_i^2's with the w_i's and the e_i's with the w_i's, respectively; $\sigma_{e^2}^2$

and σ_w^2 are the finite population variances of the e_i^2's and w_i's. In skewed populations, the correlation ρ_{ew} in (2.5) may be negligible but ρ_{e^2w} can be large and negative if units with larger $\mathbf{x}, y$ values have larger residuals but small weights. We found empirically in the simulations reported in Section 4 that ρ_{e^2w} was generally negative and larger in relative size than ρ_{ew}.

Assuming that the correlations in the last two terms of (2.5) are negligible, Spencer approximates (2.5) with

$$\text{Deff}_S \approx \left(1 - \rho_{yp}^2\right)\frac{n\bar{W}}{N} + \left(\frac{A}{\sigma_y}\right)^2\left(\frac{n\bar{W}}{N} - 1\right), \tag{2.6}$$

A similar expression is given by Park and Lee (2004; expression 4.7). Spencer proposed estimating measure (2.6) with

$$\text{deff}_S = \left(1 - R_{yp}^2\right)\text{deff}_K(\mathbf{w}) + \left(\hat{\alpha}/\hat{\sigma}_y\right)^2\left(\text{deff}_K(\mathbf{w}) - 1\right), \tag{2.7}$$

where R_{yp}^2 and $\hat{\alpha}$ are the $R-$squared and estimated intercept from fitting the model $y_i = \alpha + \beta p_i + \varepsilon_i$ with survey weighted least squares, $\hat{\sigma}_y^2 = \sum_{i\in s} w_i\left(y_i - \hat{\bar{y}}_w\right)^2 \Big/ \sum_{i\in s} w_i$ with $\hat{\bar{y}}_w = \sum_s w_i y_i \Big/ \sum_s w_i$ is the estimated population unit variance. Spencer's estimator (2.7) assumes that the population size N is large.

When ρ_{yp} is zero and σ_y is large, measure (2.7) is approximately equivalent to Kish's measure (2.4). However, Spencer's method does incorporate the survey variable y_i, unlike (2.4), and implicitly reflects the dependence of y_i on the selection probabilities p_i. We can explicitly see this by noting that when N is large, $A = \bar{Y} - BN^{-1} \approx \bar{Y}$, and (2.6) can be written as

$$\text{Deff}_S \approx \left(1 - \rho_{yp}^2\right)\frac{n\bar{W}}{N} + \frac{1}{\text{CV}_Y^2}\left(\frac{n\bar{W}}{N} - 1\right), \tag{2.8}$$

where $\text{CV}_{\bar{Y}}^2 = \sigma_y^2 \Big/ \bar{Y}^2$ is the population-level unit coefficient of variation (CV). We estimate (2.8) with

$$\text{deff}_S = \left(1 - R_{yp}^2\right)\text{deff}_K(\mathbf{w}) + \frac{1}{\text{cv}_y^2}\left(\text{deff}_K(\mathbf{w}) - 1\right), \tag{2.9}$$

where $\text{cv}_y^2 = \hat{\sigma}_y^2 \Big/ \hat{\bar{y}}_w^2$. Note that cv_y is not the standard CV produced in conventional survey estimation software, since it estimates the population unit CV of y.

3 Proposed design-effect measure

We extend Spencer's (2000) approach in single-stage sampling to produce a new weighting design effect for a calibration estimator. While Spencer's assumed $y_i = \alpha + \beta p_i + \varepsilon_i$, we model y_i as $y_i = \alpha + \mathbf{x}_i^T\boldsymbol{\beta} + \varepsilon_i = \dot{\mathbf{x}}_i^T\dot{\boldsymbol{\beta}} + \varepsilon_i$, where $\dot{\mathbf{x}}_i = [1 \quad \mathbf{x}_i]$ and $\dot{\boldsymbol{\beta}} = [\alpha \quad \boldsymbol{\beta}]$. Denote the full finite population estimators of α and $\boldsymbol{\beta}$ by $A = \bar{Y} - \bar{\mathbf{X}}\mathbf{B}$ and $\mathbf{B} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ where $\mathbf{X}$ is the $N \times p$ matrix of auxiliaries for the N units in the finite population and $\mathbf{Y}$ is the $N-$vector of y values. The finite population residuals are defined as $e_i = y_i - \left(A + \mathbf{x}_i^T\mathbf{B}\right) \equiv y_i - \dot{\mathbf{x}}_i^T\dot{\mathbf{B}}$ where $\dot{\mathbf{B}} = [A \quad \mathbf{B}]$.

Producing the design effect proposed below involves four steps: (1) constructing a linear approximation to the GREG estimator; (2) obtaining the design-variance of this linear approximation; (3) substituting model-based components into the GREG variance; and (4) taking the ratio of this model-assisted variance to the variance of the pwr−estimator of the total under srswr. Since steps (1)−(4) produce the theoretical design effect for an estimator, we add the final step: (5) plug-in sample-based estimates for each theoretical design effect component.

Step 1. A linearization of the GREG estimator (Expression 6.6.9 in Särndal et al. 1992) is

$$\begin{aligned} \hat{T}_{\text{GREG}} &\doteq \hat{T}_{\text{HT}y} + \left(\mathbf{T}_x - \hat{\mathbf{T}}_{\text{HT}x}\right)^T \dot{\mathbf{B}} \\ &= \mathbf{T}_x^T \dot{\mathbf{B}} + \sum_{i\in s} e_i / \pi_i \end{aligned} \tag{3.1}$$

where $\sum_{i\in s} e_i / \pi_i$ is the HT estimator of the population total of the e_i, $E_U = \sum_{i\in U} e_i$. To obtain a simple variance formula in step 2, we treat the case of with-replacement sampling and replace $\sum_{i\in s} e_i / \pi_i$ with the pwr−estimator $n^{-1}\sum_{i=1}^{n} e_i / p_i$. Next, define δ_i to be the number of times that unit i is selected for the sample. Since $E_\pi(\delta_i) = np_i$, the second component in (3.1) has design-expectation $E_\pi\left(n^{-1}\sum_{i=1}^{n} e_i / p_i\right) = E_U$.

Step 2. From step 1 with the assumption of with-replacement sampling, $\hat{T}_{\text{GREG}} - \mathbf{T}_x^T \dot{\mathbf{B}} \doteq n^{-1}\sum_{i=1}^{n} e_i / p_i$, with design-variance

$$\begin{aligned} \text{Var}_\pi\left(\hat{T}_{\text{GREG}} - \mathbf{T}_x^T \dot{\mathbf{B}}_U\right) &\doteq \text{Var}_\pi\left(n^{-1}\sum_{i=1}^{n} e_i / p_i\right) \\ &= n^{-1}\sum_{i\in U} p_i \left(e_i / p_i - E_U\right)^2. \end{aligned} \tag{3.2}$$

Steps 3 and 4. We follow Spencer's approach and substitute model values in variance (3.2) to formulate a design-effect measure. However, we substitute in the model-based equivalent to e_i, not y_i. Substituting the GREG residuals e_i into the variance and taking its ratio to the variance of the pwr−estimator in simple random sampling with replacement, $\text{Var}_{\text{srswr}}\left(\hat{T}_{\text{srswr}}\right) = N^2\sigma_y^2 / n$, where $\sigma_y^2 = N^{-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$, will produce our approximate design effect due to unequal calibration weighting. We can simplify things greatly by defining $u_i = A + e_i$, where $u_i = y_i - \mathbf{x}_i^T \mathbf{B}$, which implies $\bar{U} = A + \bar{E}_U = A$. The resulting design effect (see Appendix) is

$$\text{Deff}_H = \frac{n\bar{W}}{N}\left(\frac{\sigma_u^2}{\sigma_y^2}\right) + \frac{n\sigma_w}{N\sigma_y^2}\left(\rho_{u^2 w}\sigma_{u^2} - 2A\rho_{uw}\sigma_u\right) \tag{3.3}$$

where $\sigma_u^2 = N^{-1}\sum_{i=1}^{N}(u_i - \bar{U})^2, \sigma_y^2 = N^{-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2, \rho_{u^2 w}$ is the finite population correlation between u_i^2 and $w_i, \sigma_{u^2}^2$ is the variance of u_i^2 and ρ_{uw} is the correlation between u_i and w_i.

The first component in (3.3) is $O(1)$; the factor $n\bar{W}/N$ is related to the Kish deff as described below. The factor σ_u^2/σ_y^2 is an adjustment based on the effectiveness of the covariates in predicting y. The second component in (3.3) is $O(n/N)$ and incorporates terms related to the strength of the relationship between the calibration covariates and the weights.

Note that the derivation of (3.3) assumes with-replacement (WR) sampling was used. Although without replacement (WOR) sampling is more common in practice, the WOR variance of an estimated total is complicated since it involves joint selection probabilities. The WR variance formula is simple enough to provide insights into the effect of calibration on a deff. In cases where there are gains in precision from using WOR sampling, an ad hoc finite population correction factor can be incorporated in (3.3), i.e., $(1-n/N)\,\text{Deff}_H$.

Step 5. To estimate (3.3), we use

$$\text{deff}_H \approx \text{deff}_K(\mathbf{w})\frac{\hat{\sigma}_u^2}{\hat{\sigma}_y^2} + \frac{n\hat{\sigma}_w}{N\hat{\sigma}_y^2}\left(\hat{\rho}_{u^2w}\hat{\sigma}_{u^2} - 2\hat{\alpha}\hat{\rho}_{uw}\hat{\sigma}_u\right), \tag{3.4}$$

where the model parameter estimate $\hat{\alpha}$ is obtained using survey-weighted least squares, $\hat{\sigma}_y^2$ was defined in Section 2.3, $\hat{\sigma}_u^2 = \sum_{i\in s} w_i(\hat{u}_i - \bar{u}_w)^2 / \sum_{i\in s} w_i$, $\bar{\hat{u}}_w = \sum_{i\in s} w_i\hat{u}_i / \sum_{i\in s} w_i$, $\hat{u}_i = y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}_s^T\mathbf{W}\mathbf{X}_s\right)^{-1}\mathbf{X}_s^T\mathbf{W}\mathbf{y}_s$ is the survey-weighted least-squares estimate of $\boldsymbol{\beta}$, with $\mathbf{W} = \text{diag}(w_1,\ldots,w_n)$, and other terms defined in Section 2.1.

If the correlations in (3.3) are negligible or the sampling fraction n/N is small, the first term dominates and we obtain

$$\text{Deff}_H \approx \frac{n\bar{W}}{N}\left(\frac{\sigma_u^2}{\sigma_y^2}\right),$$

which can be estimated with

$$\text{deff}_H \approx \text{deff}_K(\mathbf{w})\hat{\sigma}_u^2/\hat{\sigma}_y^2. \tag{3.5}$$

Note that in samples without calibration weight adjustments, we have $\hat{u}_i = y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}} \approx y_i$ and $\sigma_u^2 \approx \sigma_y^2$. In this case expression (3.5) becomes $\text{Deff}_H \approx n\bar{W}/N$, which we estimate with Kish's measure $\text{deff}_K = 1 + [\text{CV}(\mathbf{w})]^2$. However, when the relationship between the calibration covariates $\mathbf{x}$ and y is stronger, the variance σ_u^2 should be smaller than σ_y^2. In this case, measure (3.5) is smaller than Kish's estimate. Variable weights produced from calibration adjustments are thus not as "penalized" (shown by overly high design effects) as they would be using the Kish and Spencer measures. However, if we have "ineffective" calibration, or a weak relationship between $\mathbf{x}$ and y, then σ_u^2 can be greater than σ_y^2, producing a design effect greater than one. The Spencer measure only accounts for an indirect relationship between $\mathbf{x}$ and y if there was only one x and it was used to produce p_i. This is illustrated in Section 4. We also examine the extent to which the correlation components in our proposed design effect (3.3) are large enough to influence the exact measure. Calculation of (3.3) requires only the sample $y-$values, covariates, and calibration weights. This measure can, thus, be produced more quickly than measure (2.3), whose components are often available later in data processing after a variance estimation system is set up.

4 Empirical evaluation

We conducted two simulation studies using data that mimic single-stage sampling. The first utilizes publically-available data from tax returns and continuous variables of interest, while the second examines the performance of the alternative measures for a binary outcome measure in a single-stage survey.

4.1 Establishment data simulation study

Here a sample dataset of tax return data is used to mimic an establishment survey setup. The data come from the Tax Year 2007 Statistics of Income (SOI) Form 990 Exempt Organization (EO) sample. This is a stratified Bernoulli sample of 22,430 EO tax returns selected from 428,719 filed with and processed by the IRS between December 2007 and November 2009. This sample dataset, along with the population frame data, is free and electronically available online (Statistics of Income 2011). These data make a candidate "establishment-type" dataset for estimating design effects, in which Kish's design effect may not apply.

The SOI EO sample dataset is used here as a pseudopopulation for illustration. Four variables of interest are used: Total Assets, Total Liabilities, Total Revenue, and Total Expenses. Returns that were sampled with certainty or that had "very small" assets (defined by having Total Assets less than $1,000,000, including zero) were removed, leaving 8,914 units. We then randomly replicated and perturbed the data to create a pseudopopulation of 50,000 units. We used simple random sampling with replacement to select more observations, then the additional data values were perturbed using the `jitter` (Chambers, Cleveland, Kleiner and Tukey 1983) function in R.

Figure 4.1 shows a pairwise plot of the pseudo-population, including plots of the variable values against each other in the lower left panels, histograms on the diagonal panels, and the correlations among the variables in the upper right panels. This plot mimics establishment-type data patterns. From the diagonal panels, we see that the variables of interest are all highly skewed. From the lower left panels, there exists a range of different relationships among them. The Total Assets variable is less related to Total Revenue and Total Expenses (with moderate correlations of $0.41-0.44$); Total Revenue and Total Expenses are highly correlated.

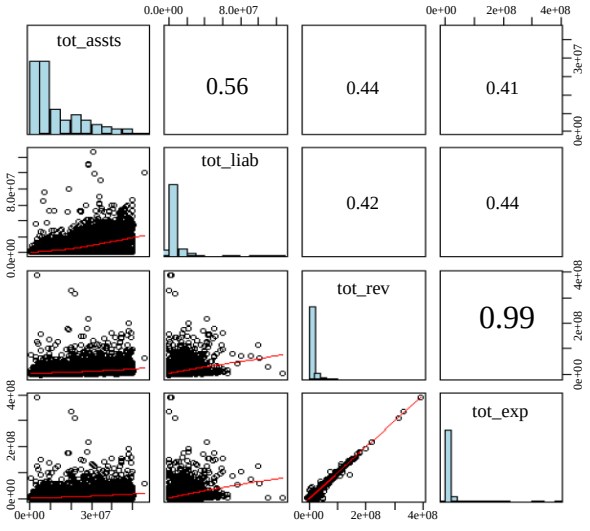

Figure 4.1 Pseudopopulation values and loess lines for design effect evaluation.

Three sizes of samples were selected $(n = 100; 500; 1{,}000)$ without replacement from the pseudopopulation using the square root of Total Assets as a measure of size. This type of sampling is referred to as πps sampling subsequently. The HT weights were then calibrated using the "`linear`" method in the `calibrate` function in the `survey` package for R (corresponding to a GREG estimator, Lumley 2012) to match the totals of an intercept, Total Assets and Total Revenue. The analysis variables are Total Liabilities and Total Expenses. (Note that we follow the common practice of developing procedures in the previous sections using formulas for with-replacement sampling but empirically evaluating them in without-replacement samples, which are the type used in applications.)

Eight design effects estimates are considered:

- Estimates of the design effect measures (2.2) and (2.3). Expression (2.2) reflects the efficiency of πps sampling and use of the HT$-$estimator. Expression (2.3) reflects gains (if any) of πps sampling combined with GREG estimation;
- The Kish measure (2.4) computed using the GREG weights;
- Three Spencer measures computed using the GREG weights: (i) the exact measure that estimates (2.5), (ii) the approximation (2.7) assuming zero correlation terms, and (iii) the large-population approximation (2.9). The Spencer measures are designed to reflect gains due to PPSWR sampling and use of the pwr$-$estimator. It does not account for any gains due to calibration.
- Two proposed measures: (i) the exact proposed single-stage design effect (3.4) and (ii) the zero-correlation approximation (3.5). Both of these are meant to show the precision gains (if any) of PPSWR sampling combined with GREG estimation.

Note that neither the Spencer nor the proposed measures account for any reduction in variances due to sampling a large fraction of the population.

We selected ten thousand samples to further understand the empirical behavior of the alternative design effect estimators. The empirical relbiases and ratio of the mean square errors (MSE's) of the totals are

$$\begin{aligned}\text{relbias}\left(\hat{T}\right) &= 100 \times \sum_{s=1}^{S}\left(\hat{T}_s - T\right)\Big/T \\ \text{MSE ratio} &= \text{MSE}\left(\hat{T}_{\text{HT}}\right)\Big/\text{MSE}\left(\hat{T}_{\text{GREG}}\right) \\ &= \sum_{s=1}^{S}\left(\hat{T}_{\text{HT},s} - T\right)^2 \Big/ \sum_{s=1}^{S}\left(\hat{T}_{\text{GREG},s} - T\right)^2\end{aligned}$$

where $\hat{T}_s$ is an estimated total from sample s (either HT or GREG), $S = 10{,}000$ is the number of samples selected, and $\hat{T}_{\text{HT},s}$ and $\hat{T}_{\text{GREG},s}$ are the estimated HT and GREG totals from sample s. The empirical deff of an estimated total is computed as $\text{empdeff}\left(\hat{T}\right) = S^{-1}\sum_{s=1}^{S}\left(\hat{T}_s - \bar{\hat{T}}\right)^2 \Big/ \text{Var}_{\text{srswr}}\left(\hat{T}_{\text{srswr}}\right)$ where $\bar{\hat{T}} = S^{-1}\sum_{s=1}^{S}\hat{T}_s$ and $\text{Var}_{\text{srswr}}\left(\hat{T}_{\text{srswr}}\right) \doteq N^2\sigma_y^2/n$.

The results for relbiases and MSEs are shown in Table 4.1. Both estimators of totals are approximately unbiased. The GREG is also more precise than the HT estimator, especially for Total Expenses, as evidenced by the MSE ratios larger than one.

Table 4.1
Simulation results of HT and GREG totals, 10,000 πps samples drawn from the SOI 2007 pseudopopulation EO data

	Variable of Interest					
	Total Liabilities (weakly correlated with X)			Total Expenses (strongly correlated with X)		
Estimates	$n = 100$	$n = 500$	$n = 1{,}000$	$n = 100$	$n = 500$	$n = 1{,}000$
Percent relbias(HT)	-0.13	0.07	0.03	-0.64	0.05	0.07
Percent relbias(GREG)	0.37	0.27	0.14	-0.12	-0.01	0.00
MSE ratio	1.17	1.20	1.19	34.89	50.11	48.26

Note A small number of samples were dropped in which either the matrix to be inverted for the GREG was singular or the GREG produced negative weights. The percentages of samples dropped were 3.6% for $n = 100$, 1.2% for $n = 500$, and 0.5% for $n = 1{,}000$.

We also computed the biases of the various estimated design effects across the 10,000 samples. The relbiases of the Kish, Spencer, and proposed design effect estimates are computed as

$$\text{relbias}\left(\text{deff}_K\right) = 100 \times \left(\overline{\text{deff}}_K - \text{edeff}\left(\hat{T}_{\text{HT}y}\right)\right) \Big/ \text{edeff}\left(\hat{T}_{\text{HT}y}\right),$$

$$\text{relbias}\left(\text{deff}_S\right) = 100 \times \left(\overline{\text{deff}}_S - \text{edeff}\left(\hat{T}_{\text{HT}y}\right)\right) \Big/ \text{edeff}\left(\hat{T}_{\text{HT}y}\right),$$

and

$$\text{relbias}\left(\text{deff}_H\right) = 100 \times \left(\overline{\text{deff}}_H - \text{edeff}\left(\hat{T}_{\text{GREG}}\right)\right) \Big/ \text{edeff}\left(\hat{T}_{\text{GREG}}\right)$$

where $\overline{\text{deff}}_K$, $\overline{\text{deff}}_S$, and $\overline{\text{deff}}_H$ are the average Kish, Spencer, and proposed deff's over all samples. The terms $\text{edeff}\left(\hat{T}_{\text{HT}y}\right)$ and $\text{edeff}\left(\hat{T}_{\text{GREG}}\right)$ are computed in two ways: (1) as the simulation empdeff of $\hat{T}_{\text{HT}y}$ $\left(\text{or } \hat{T}_{\text{GREG}}\right)$, and (2) as the average over all samples of the deff's of $\hat{T}_{\text{HT}y}$ computed from the `survey` package. The `survey` package's default method of estimating the deff from a particular sample uses a with-replacement variance estimate in the numerator. This corresponds to the sample design used to derive deff_H. Results are displayed in Table 4.2.

For both variables of interest, we see large positive biases for the Kish design effect, and the design effects involving approximations. Thus, ignoring correlation components accounted for in the 'exact' Spencer and proposed design effects would lead to over-estimating the design effects.

The proposed estimator is closer to the `survey` package design effects than to the empirical simulation deff's of the GREG. Although the relbiases of deff_H are fairly large for Total Expenses when computed with respect to edeff, the empirical deff's themselves are small. We highlight the small magnitude of the Total Expenses (y_2) variable deff of 0.02 to put the relbiases into context. For example, the relbias of 12.9% for the exact version of our proposed estimator for $n = 500$ for y_2 corresponds to a difference in the third decimal place. Specifically, in this scenario, on average we over-estimate the deff by 0.003.

We can understand why calibration is more efficient for Expenses than for Liabilities by examining the distributions of y_i and u_i in one particular sample. Figures 4.2 and 4.3 show boxplots of u_i and y_i for each variable and sample size.

Table 4.2
Relative bias of design effect estimates, 10,000 πps samples drawn from the SOI 2007 pseudopopulation EO data

	Variable of Interest					
	Total Liabilities (weakly correlated with X)			Total Expenses (strongly correlated with X)		
	$n = 100$	$n = 500$	$n = 1{,}000$	$n = 100$	$n = 500$	$n = 1{,}000$
*Empirical deff's**						
HT	0.51	0.50	0.50	0.56	0.65	0.64
GREG	0.43	0.42	0.42	0.02	0.02	0.02
Relative biases w.r.t. empirical deff's						
*Kish***	158.7	158.3	158.3	132.8	101.7	104.7
*Spencer***						
Exact	2.6	2.0	1.8	9.9	-4.5	-2.2
Zero-corr. approx.	96.1	98.0	98.4	91.2	70.1	73.7
Large $-N$ approx.	96.7	98.9	99.3	101.7	78.1	81.7
*Proposed****						
Exact	-6.3	-1.6	0.2	25.3	12.9	8.1
Zero-corr. approx.	83.4	94.0	98.2	129.9	116.6	108.7
Relative biases w.r.t. average of `survey` *package deff's*						
*Kish***	219.7	211.3	209.4	6,400.5	7,786.2	8,287.2
*Spencer***						
Exact	3.1	0.8	0.5	3.5	-1.0	-1.5
Zero-corr. approx.	97.1	95.8	95.8	80.1	76.2	74.8
Large $-N$ approx.	97.7	96.7	96.7	90.0	84.5	82.8
*Proposed****						
Exact	-0.9	-0.2	-0.1	11.3	-0.4	-0.1
Zero-corr. approx.	94.0	96.8	97.6	104.2	91.0	93.0

* Averages across the simulated samples;
** relative to the average of empirical HT deff's;
*** relative to the average of empirical GREG deff's.

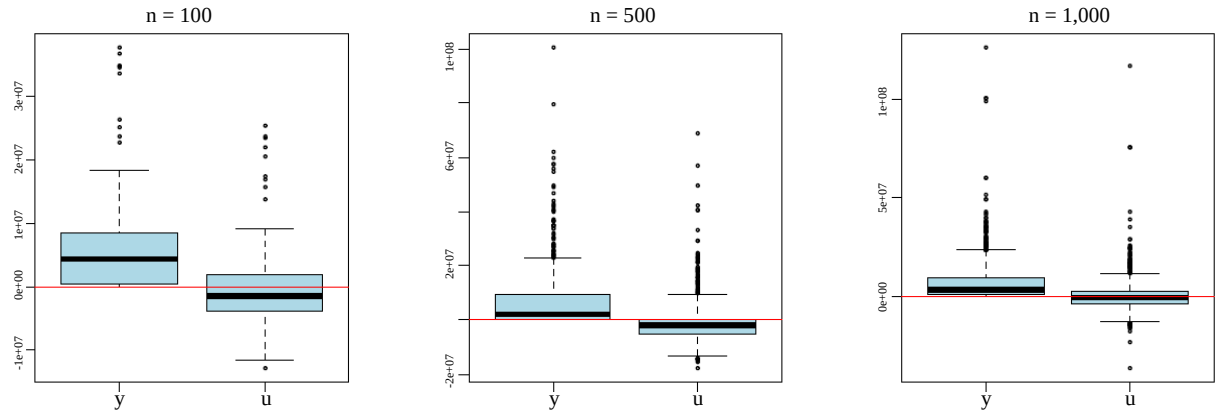

Figure 4.2 Boxplots of y_i and u_i-values from ppswr samples from the 2007 SOI EO data, total liabilities variable (weakly correlated with X).

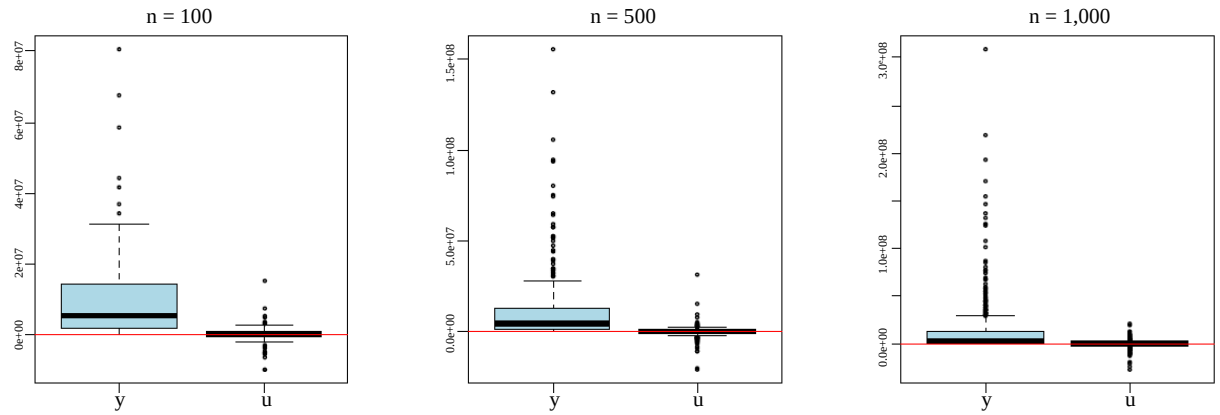

Figure 4.3 Boxplots of y_i and u_i-values from πps samples from the 2007 SOI EO data, total expenses variable (strongly correlated with X).

The u_i-values in all of these samples have shorter ranges of values and less variation than y_i, particularly for the Total Expenses variable. This occurs since the Total Expenses variable is highly correlated with the calibration variable Total Revenue (see Figure 4.1) and explains why the direct and proposed design effect measures are so much smaller for Total Expenses.

4.2 Simulation study with a binary variable

The second simulation study illustrates the performance of the proposed estimator when estimating the total of a binary variable in a single-stage survey that uses poststratification.

We use the `nhis.large` population, which has $N = 21{,}588$ units, from the `PracTools` R package (Valliant, Dever and Kreuter 2015) to gauge the impact of poststratification weighting adjustments. The binary variable used is whether or not a person received Medicaid or not. Receipt of Medicaid, which is a social welfare program in the US, is an example of a variable that is collected in some telephone surveys. Missing values of Medicaid recipiency were recoded to be "no" responses. There is a fairly strong relationship between race-ethnicity, age, and whether Medicaid is received, as shown in Table 4.3 or Table 14.1 in Valliant, Dever and Kreuter (2013). The 15 age × race-ethnicity cells in the table will be used as poststrata, which is a typical procedure in telephone surveys.

Table 4.3
Population percentages of persons receiving medicaid, by age group and Hispanic status

	Hispanic Status		
Age Group	**Hispanic**	**Non-Hispanic White**	**Non-Hispanic Black or Other**
< 18 years	31.8	12.9	30.9
18-24	10.5	6.5	12.2
25-44	7.5	3.8	8.6
45-64	2.4	3.0	6.2
65+	26.8	3.7	16.2

In our simulation, we selected 10,000 simple random samples without replacement from the NHIS population. The HT estimator for the total number of persons receiving Medicaid is $N\bar{y}_s$, where $\bar{y}_s$ is the proportion in sample s that receives Medicaid. Due to the relatively large number of poststrata and varying number of persons receiving Medicaid by poststratum, we include results only for samples of size $n = 500$ and $1{,}000$ since no collapsing of poststrata within a given particular sample was needed for these sample sizes.

The base weights for the HT−estimator are simply $w_i = N/n$. The variance of the poststratified estimator is 91% of that of $N\bar{y}_s$ in samples of $n = 500$ and 88% in samples of $n = 1{,}000$. Since the base weights are constant, Spencer's design effects are not computable in this example. Therefore, only results for the Kish and proposed design effects are shown in Table 4.4.

Table 4.4
Relative bias of design effect estimates, 10,000 pps samples drawn from the NHIS pseudopopulation data

	Number of Persons Receiving Medicaid			
	$n = 500$		$n = 1{,}000$	
*Empirical deff's**				
HT	0.97		0.95	
GREG	0.91		0.88	
	w.r.t. empirical deff	w.r.t. survey deff	w.r.t. empirical deff	w.r.t. survey deff
Relative biases (percent)				
*Kish***	6.0	17.5	7.0	17.6
*Proposed****				
Exact	-1.4	3.2	-0.9	5.0
Zero-corr. approx.	-1.5	2.9	-1.2	4.7

* Averages across the simulated samples;
** relative to the average of empirical HT deff's;
*** relative to the average of empirical GREG deff's.

The Kish design effect has positive biases of 17.5% and 17.6% when computed with respect to the empirical deff's. The exact proposed design effects are positively biased with respect to the survey deff (3.2 and 5.0%), but much less so than the Kish estimator. In this example, the zero-correlation approximation is very similar to the exact version of the proposed estimator. The correlation components were negligible for these weighting adjustments within three decimal places.

5 Discussion, limitations, and conclusions

We propose a new design effect that gauges the impact of calibration weighting adjustments on an estimated total in single-stage sampling. Two existing design effects are the Kish (1965) "design effect due to weighting" and one due to Spencer (2000). Both of these are inadequate to reflect efficiency gains due to calibration. The Kish deff is a reasonable measure if equal weighting is optimal or nearly so, but does not reveal efficiencies that may accrue from sampling with varying probabilities. The Spencer deff

does signal whether the HT (or pwr) estimator in varying probability sampling is more efficient than srs. But, the Spencer deff does not reflect any gains from using calibration.

The proposed design effect measures the impact of both sampling with varying probabilities and of using a calibration estimator, like the GREG, that takes advantage of auxiliary information. As we demonstrate empirically, the proposed design effects do not penalize unequal weights when the relationship between the survey variable and calibration covariate is strong. We also demonstrated empirically that the correlation components in the Spencer measure and our proposed measure can be important in some situations. It is not overly difficult to calculate these components, and these should be incorporated when possible to avoid over estimates of the design effects. However, the high correlations between survey and auxiliary variables that we observed in our establishment pseudopopulation data may be unattainable for some surveys that lack auxiliary information. In cases where the auxiliary information is ineffective or is not used, the proposed measure approximates Kish's deff. The measure presented here is applicable to single-stage sampling but can be extended to more complex sample designs, like cluster sampling.

Our measure uses the model underlying the general regression estimator to extend the Spencer measure. The survey variable, covariates, and weights are required to produce the design effect estimate. Since the variance (3.2) is approximately correct in large samples for all calibration estimators, our design effect should reflect the effects of many forms of commonly used weighting adjustment methods, including poststratification, raking, and the GREG estimator. Although design effects that do account for these adjustments can be computed directly from estimated variances, it is important for practitioners to understand that the existing Kish and Spencer deff's do not reflect any gains from those adjustments. The deff introduced in this paper, thus, serves as a corrective to that deficiency.

For practical consideration, the deff in (3.4) is available in the `deffH` function in the R `PracTools` package; see Valliant et al. (2015) for documentation and examples.

Acknowledgements

We thank the referees for their thorough reviews which improved the presentation. Any opinions expressed are those of the authors and do not reflect those of the Internal Revenue Service.

Appendix

Proposed design effect in single-stage sampling

The appendix sketches the derivation of the proposed deff. Most notation was defined in the previous sections of the paper. The average population one-draw probability is $\bar{P} = N^{-1}\sum_{i=1}^{N} p_i$. Assume that the design satisfies $\bar{P} = N^{-1}$. Consider the model $y_i = \alpha + \mathbf{x}_i^T\boldsymbol{\beta} + \varepsilon_i$. If the full finite population were available, then the least-squares population regression line would be

$$y_i = A + \mathbf{x}_i^T\mathbf{B} + e_i, \tag{A.1}$$

where A and $\mathbf{B}$ are the values found by fitting an ordinary least squares regression line in the full finite population. That is, $A = \bar{Y} - \mathbf{B}\bar{\mathbf{X}}$, $\mathbf{B} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$, where $\mathbf{X}$ is the $N \times p$ population matrix of auxiliary variables, $\bar{Y} = N^{-1}\sum_{i=1}^{N} y_i$ is the population mean, and $\bar{\mathbf{X}}$ is the vector of population means of the x's. The e_i's are defined as the finite population residuals, $e_i = y_i - A - \mathbf{x}_i^T\mathbf{B}$, and are not superpopulation model errors. Denote the population variance of the y's, e's, e^2, and weights as $\sigma_y^2, \sigma_e^2, \sigma_{e^2}^2, \sigma_w^2$, e.g., $\sigma_y^2 = N^{-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$, and the finite population correlations between the variables in the subscripts as ρ_{yp}, ρ_{ew}, and ρ_{e^2w}. The GREG theoretical design-variance in with-replacement sampling is

$$\begin{aligned} \mathrm{Var}\left(\hat{T}_{\mathrm{GREG}}\right) &= n^{-1}\sum_{i=1}^{N} p_i\left(e_i/p_i - E_U\right)^2 \\ &= n^{-1}\left(\sum_{i=1}^{N} e_i^2/p_i - E_U^2\right), \end{aligned} \tag{A.2}$$

where $E_U = \sum_{i=1}^{N} e_i$. Using the model in (A.1) produces a design effect with several complex terms, many of which contain correlations that cannot be dropped as in Spencer's approximation. The design effect can be simplified using an alternative formulation: $u_i = A + e_i$, where $u_i = y_i - \mathbf{x}_i^T\mathbf{B}$. First, we rewrite the population total of the e_i's as $E_U = \sum_{i=1}^{N} e_i = N\bar{U} - NA$, where $\bar{U} = N^{-1}\sum_{i=1}^{N} u_i$. From this, $E_U^2 = (N\bar{U})^2 + (NA)^2 - 2N^2\bar{U}A$. Second, using $w_i = (np_i)^{-1}$, or $p_i = (nw_i)^{-1}$, we rewrite the component $\sum_{i=1}^{N} e_i^2/p_i$ as

$$\begin{aligned} \sum_{i=1}^{N} e_i^2/p_i &= \sum_{i=1}^{N} \frac{(u_i - A)^2}{(nw_i)^{-1}} \\ &= n\sum_{i=1}^{N} w_i u_i^2 + nA^2\sum_{i=1}^{N} w_i - 2nA\sum_{i=1}^{N} w_i u_i. \end{aligned} \tag{A.3}$$

Subtracting E_U^2 from (A.3) and dividing by n gives

$$\begin{aligned} n^{-1}\left(\sum_{i=1}^{N} e_i^2/p_i - E_U^2\right) &= \sum_{i=1}^{N} w_i u_i^2 - n^{-1}(N\bar{U})^2 \\ &\quad + A^2\left(\sum_{i=1}^{N} w_i - n^{-1}N^2\right) \\ &\quad + n^{-1}2N^2\bar{U}A - 2A\sum_{i=1}^{N} w_i u_i. \end{aligned} \tag{A.4}$$

Following Spencer's approach using the covariance substitutions, the first and fifth terms in (A.4) can be rewritten as $\sum_{i=1}^{N} w_i u_i^2 = N\rho_{u^2w}\sigma_{u^2}\sigma_w + N\bar{W}\left(\sigma_u^2 + \bar{U}^2\right)$ and $\sum_{i=1}^{N} w_i u_i = N\rho_{uw}\sigma_u\sigma_w + N\bar{W}\bar{U}$. Plugging these back into the variance (A.4) gives

$$\begin{aligned} n^{-1}\left(\sum_{i=1}^{N} e_i^2/p_i - E_U^2\right) &= N\rho_{u^2w}\sigma_{u^2}\sigma_w + N\bar{W}\left(\sigma_u^2 + \bar{U}^2\right) - n^{-1}(N\bar{U})^2 \\ &\quad + NA^2\left(\bar{W} - n^{-1}N\right) \\ &\quad + 2n^{-1}N^2\bar{U}A - 2A\left(N\rho_{uw}\sigma_u\sigma_w + N\bar{W}\bar{U}\right). \end{aligned} \tag{A.5}$$

The variance of the pwr−estimator under simple random sampling with replacement, where $p_i = N^{-1}$, reduces to $\text{Var}_{\text{srswr}}\left(\hat{T}_{\text{pwr}}\right) = N^2\sigma_y^2/n$. Taking the ratio of (A.5) to the pwr−variance gives the following design effect:

$$\begin{aligned}\text{Deff}_H &= \text{Var}_{\text{GREG}}\left(\hat{T}_{\text{cal}}\right)/\text{Var}_{\text{srswr}}\left(\hat{T}_{\text{pwr}}\right) \\ &= \frac{n\bar{W}}{N}\left(\frac{\sigma_u^2}{\sigma_y^2}\right) + \frac{\left(\bar{U} - A\right)^2}{\sigma_y^2}\left(\frac{n\bar{W}}{N} - 1\right) \\ &+ \frac{n\sigma_w}{N\sigma_y^2}\left(\rho_{u^2w}\sigma_{u^2} - 2A\rho_{uw}\sigma_u\right).\end{aligned} \tag{A.6}$$

Since $u_i = A + e_i$, $\bar{U} = A$, (A.6) becomes

$$\text{Deff}_H = \frac{n\bar{W}}{N}\left(\frac{\sigma_u^2}{\sigma_y^2}\right) + \frac{n\sigma_w}{N\sigma_y^2}\left(\rho_{u^2w}\sigma_{u^2} - 2A\rho_{uw}\sigma_u\right). \tag{A.7}$$

We estimate measure (A.7) with

$$\text{deff}_H \approx \left(1 + \left[\text{CV}\left(\mathbf{w}\right)\right]^2\right)\frac{\hat{\sigma}_u^2}{\hat{\sigma}_y^2} + \frac{n\hat{\sigma}_w}{N\hat{\sigma}_y^2}\left(\hat{\rho}_{u^2w}\hat{\sigma}_{u^2} - 2\hat{\alpha}\hat{\rho}_{uw}\hat{\sigma}_u\right), \tag{A.8}$$

where the model parameter estimates are defined in Sections 2.3 and 3.

References

Brick, M., and Montaquila, J. (2009). Nonresponse. In *Handbook of Statistics, Sample Surveys: Design, Methods and Application*, (Eds., D. Pfeffermann and C.R. Rao), 29A, Amsterdam: Elsevier BV.

Chambers, J.M., Cleveland, W.S., Kleiner, B. and Tukey, P.A. (1983). *Graphical Methods for Data Analysis*. Pacific Grove CA: Wadsworth.

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

Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

Hansen, M.H., and Hurwitz, W.N. (1943). On the theory of sampling from a finite population. *Annals of Mathematical Statistics*, 14, 333-362.

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

Kalton, G., and Flores-Cervantes, A. (2003). Weighting methods. *Journal of Official Statistics*, 19 (2), 81-97.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Kish, L. (1990). Weighting: Why, when, and how? *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods,* American Statistical Association, 121-129.

Kish, L. (1992). Weighting for unequal Pi. *Journal of Official Statistics,* 8, 183-200.

Kott, P. (2009). Calibration weighting: Combining probability samples and linear prediction models. In *Handbook of Statistics, Sample Surveys: Design, Methods and Application,* (Eds., D. Pfeffermann and C.R. Rao), 29B, Amsterdam: Elsevier BV.

Lumley, T. (2012). Survey: Analysis of complex survey samples. R package version 3.28-2.

Park, I., and Lee, H. (2004). Design effects for the weighted mean and total estimators under complex survey sampling. *Survey Methodology*, 30, 2, 183-193.

Rao, J.N.K., and Scott, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.

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

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

Spencer, B.D. (2000). An approximate design effect for unequal weighting when measurements may correlate with selection probabilities. *Survey Methodology,* 26, 2, 137-138.

Statistics of Income (2011). 2007 Charities & Tax-Exempt Microdata Files. Available at: http://www.irs.gov/uac/SOI-Tax-Stats-2007-Charities-&-Tax-Exempt-Microdata-Files.

Valliant, R., Dever, J.A. and Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer.

Valliant, R., Dever, J.A. and Kreuter, F. (2015). PracTools: Tools for Designing and Weighting Survey Samples. R package version 0.2. http://CRAN.R-project.org/package=PracTools.