A Design Effect Measure for Calibration Weights in Single-Stage Samples

Kimberly Henry[1] and Richard Valliant[2]
[1]Statistics of Income, Internal Revenue Service
77 K Street, NE, Washington, DC 20002
[2] Universities of Michigan & Maryland, College Park, MD 20854.

Abstract

We propose a model-based extension of weighting design-effect measures. We develop a summary-level diagnostic 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 using a case study and simulation involving establishment-type data.

Key words: Auxiliary data; Kish weighting design effect; Spencer design effect; generalized regression estimator

1. Introduction

The most popular measure to summarize the impact of differential weights is Kish's (1965, 1992) design effect due to weighting. Spencer (2000) proposed a simple model-based approach that depends on a single covariate to estimate the impact on variance of using variable weights in a situation where the analysis variable depends on the single covariate.

However, these approaches 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 Kish design effects attempt to measure the impact of variable weights, they are informative only under special circumstances, do not account for alternative variables of interest, and can incorrectly measure the impact of differential weighting in some circumstances (Kish 1992). 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.

In particular, the Kish and Spencer measures, reviewed in section 2, may not accurately produce design effects for unequal weighting induced by calibration adjustments. These 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 Cervantes-Flores 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.

We extend these existing design effects to produce a new measure that summarizes the impact of calibration weight adjustments before and after they are applied to single-stage survey weights. The proposed measure in section 3 accounts for the joint effect of a non-*epsem* sample design and unequal weight adjustments in the larger class of calibration estimators. Our summary measure incorporates the survey variable like Spencer's model, using a generalized regression variance to reflect multiple calibration covariates. In section 4, we apply the estimators in a case study and simulation involving establishment-type survey data and demonstrate empirically how the proposed estimator outperforms the existing methods in the presence of unequal calibration weights.

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)*:

$$\hat{T}_{GREG} = \hat{T}_{HTy} + \hat{\mathbf{B}}^T\left(\mathbf{T}_x - \hat{\mathbf{T}}_{HTx}\right) = \sum\nolimits_{i\in s} g_i y_i / \pi_i \;, \tag{1}$$

where $\hat{T}_{HTy} = \sum\nolimits_{i\in s} y_i/\pi_i$ is the Horvitz-Thompson (HT, 1952) estimator of the population total of y, $\hat{\mathbf{T}}_{HTx} = \sum\nolimits_{i\in s} \mathbf{x}_i/\pi_i$ is the vector of HT estimated totals for the auxiliary variables, $\mathbf{T}_x = \sum\nolimits_{i=1}^{N} \mathbf{x}_i$ is the corresponding vector of known totals, $\hat{\mathbf{B}} = \mathbf{A}_s^{-1}\mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathbf{\Pi}_s^{-1}\mathbf{y}_s$ is the regression coefficient, with $\mathbf{A}_s = \mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathbf{\Pi}_s^{-1}\mathbf{X}_s$, $\mathbf{X}_s^T$ is the matrix of $\mathbf{x}_i$ values in the sample, $\mathbf{V}_{ss} = diag\left(v_i\right)$ is the diagonal of the variance matrix specified under the working model (defined below), and $\mathbf{\Pi}_s = diag\left(\pi_i\right)$. In the second expression for the GREG estimator in (1), $g_i = 1 + \left(\mathbf{T}_x - \hat{\mathbf{T}}_{HTx}\right)^T \mathbf{A}_s^{-1}\mathbf{x}_i v_i^{-1}$ 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, $y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \varepsilon_i$, $\varepsilon_i \sim \left(0, v_i\right)$. The GREG is consistent and approximately design-unbiased when the sample size is large (Deville and Särndal 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 (Särndal *et. al* 1992; Särndal *et. al* 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

$$Deff\left(\hat{T}\right)=Var_{\pi}\left(\hat{T}\right)\Big/Var_{srswr}\left(\hat{T}_{HT}\right). \tag{2}$$

We refer to this as a "direct" estimator because it uses theoretical variances in the numerator and denominator. The alternatives that are presented subsequently use various approximations to the components in (2). The design effect in (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 simple random sample with replacement (*srswr*) of the same size had been used.

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

$$Deff\left(\hat{T}_{cal}\right)=Var_{GREG}\left(\hat{T}_{cal}\right)\Big/Var_{srswr}\left(\hat{T}_{HT}\right). \tag{3}$$

To estimate these design-effects, we use the appropriate corresponding sample-based variance estimates. Estimates of both measures (2) and (3) can be produced using conventional survey estimation software. Our proposed design effect is a model-based approximation to (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 $\mathbf{w}=\left(w_1,\ldots,w_n\right)^T$, this measure is

$$\begin{aligned} deff_K\left(\mathbf{w}\right) &= 1+\left[CV\left(\mathbf{w}\right)\right]^2 \\ &= \frac{n\sum_{i\in s}w_i^2}{\left[\sum_{i\in s}w_i\right]^2} \quad , \end{aligned} \tag{4}$$

where $CV\left(\mathbf{w}\right)=\sqrt{n^{-1}\sum_{i\in s}\left(w_i-\bar{w}\right)^2\Big/\bar{w}^2}$ is the coefficient of variation of the weights with $\bar{w}=n^{-1}\sum_{i\in s}w_i$. Expression (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.

2.4. Spencer's Model-based 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). Suppose that p_i is the one-draw probability of selecting unit i, which is correlated with y_i and that a linear model holds for y_i: $y_i=\alpha+\beta p_i+\varepsilon_i$. A particular case of this would be $p_i \propto x_i$, where x_i is a measure of size associated with unit 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 $B=\sum_{i\in U}\left(y_i-\bar{Y}\right)\left(p_i-\bar{P}\right)\Big/\sum_{i\in U}\left(p_i-\bar{P}\right)^2$, where $\bar{Y},\bar{P}$ are the finite population means for y_i and p_i. The finite population variance of the residuals, $e_i=y_i-\left(A+Bp_i\right)$, is

$\sigma_e^2 = \left(1-\rho_{yp}^2\right) N^{-1} \sum_{i \in U} \left(y_i - \bar{Y}\right)^2 = \left(1-\rho_{yp}^2\right)\sigma_y^2$, where ρ_{yp} is the finite population correlation between y_i and p_i. The usual base weight under *PPSWR*-sampling is $w_i = \left(np_i\right)^{-1}$. The estimated total studied by Spencer is referred to as the *pwr*-estimator (Särndal *et al.* 1992) and is defined as $\hat{T}_{pwr} = \sum_{i \in s} w_i y_i$, with design-variance $Var\left(\hat{T}_{pwr}\right) = n^{-1} \sum_{i \in U} p_i \left(y_i / p_i - T\right)^2$ in single-stage sampling. 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):

$$Deff_S = \frac{A^2}{\sigma_y^2}\left(\frac{n\bar{W}}{N}-1\right) + \frac{n\bar{W}}{N}\left(1-\rho_{yp}^2\right) + \frac{n\rho_{e^2 w}\sigma_{e^2}\sigma_w}{N\sigma_y^2} + \frac{2An\rho_{ew}\sigma_e\sigma_w}{N\sigma_y^2}. \tag{5}$$

Assuming that the correlations in the last two terms of (5) are negligible, Spencer approximates (5) with

$$Deff_S \approx \left(1-\rho_{yp}^2\right)\frac{n\bar{W}}{N} + \left(\frac{A}{\sigma_y}\right)^2 \left(\frac{n\bar{W}}{N}-1\right), \tag{6}$$

where $\bar{W} = N^{-1} \sum_{i \in U} w_i = \left(nN\right)^{-1} \sum_{i \in U} 1/p_i$ is the average weight in the population. Spencer proposed estimating measure (6) with

$$deff_S = \left(1-R_{yp}^2\right) deff_K\left(\mathbf{w}\right) + \left(\hat{\alpha}/\hat{\sigma}_y\right)^2 \left(deff_K\left(\mathbf{w}\right)-1\right), \tag{7}$$

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 (7) has a large-*N* approximation assumption.

When ρ_{yp} is zero and σ_y is large, measure (7) is approximately equivalent to Kish's measure (4). However, Spencer's method does incorporate the survey variable y_i, unlike (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 (6) can be written as

$$Deff_S \approx \left(1-\rho_{yp}^2\right)\frac{n\bar{W}}{N} + \frac{1}{CV_Y^2}\left(\frac{n\bar{W}}{N}-1\right), \tag{8}$$

where $CV_Y^2 = \sigma_y^2 \big/ \bar{Y}^2$ is the population-level unit coefficient of variation (CV). We estimate (8) with

$$deff_S = \left(1-R_{yp}^2\right) deff_K\left(\mathbf{w}\right) + \frac{1}{cv_y^2}\left(1-deff_K\left(\mathbf{w}\right)\right), \tag{9}$$

where $cv_y^2 = \hat{\sigma}_y^2 \big/ \hat{\bar{y}}_w^2$. 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

Here we extend Spencer's (2000) approach in single-stage sampling to produce a new weighting design effect measure for a calibration estimator. While Spencer's assumed $y_i = \alpha + \beta p_i + \varepsilon_i$, here 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 $\dot{\mathbf{x}}_i = \begin{bmatrix} 1 & \mathbf{x}_i \end{bmatrix}$ and $\dot{\boldsymbol{\beta}} = \begin{bmatrix} \alpha & \boldsymbol{\beta} \end{bmatrix}$. Denote the full finite population estimators of α and $\boldsymbol{\beta}$ by A and $\mathbf{B}$ and the finite population residuals 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 $\dot{\mathbf{B}} = \begin{bmatrix} A & \mathbf{B} \end{bmatrix}$.

We produce the design effect in four steps: (1) constructing a linear approximation to the GREG estimator; (2) obtaining the variance of this linear approximation; (3) substituting our model-based components into the GREG variance; and (4) taking the ratio of the model-based 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: (5) plug-in sample-based estimates for each theoretical design effect component.

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

$$\begin{aligned} \hat{T}_{GREG} &\doteq \hat{T}_{HTy} + \left(\mathbf{T}_x - \hat{\mathbf{T}}_{HTx} \right)^T \dot{\mathbf{B}} \\ &= \mathbf{T}_x^T \dot{\mathbf{B}} + \tilde{e}_U \\ &\doteq \mathbf{T}_x^T \dot{\mathbf{B}} + n^{-1} \sum\nolimits_{i \in s} e_i / p_i \end{aligned} \qquad (10)$$

where $\tilde{e}_U = \sum\nolimits_{i \in s} e_i / \pi_i$ is the HT estimator of the population total of the e_i, $E_U = \sum\nolimits_{i \in U} e_i$. The last line of (10) holds if we assume that with-replacement sampling was used and that p_i is small enough that $\pi_i \doteq n p_i$. Next, define δ_i to be the number of times that unit i is selected for the sample. Since $E_\pi \left(\delta_i \right) = n p_i$, the second component in (10) has design-expectation $E_\pi \left(n^{-1} \sum\nolimits_{i \in s} e_i / p_i \right) = E_U$.

Step 2. From (10), $\hat{T}_{GREG} - \mathbf{T}_x^T \dot{\mathbf{B}} \doteq n^{-1} \sum\nolimits_{i \in s} e_i / p_i$, with design-variance

$$\begin{aligned} Var_\pi \left(\hat{T}_{GREG} - \mathbf{T}_x^T \dot{\mathbf{B}}_U \right) &= Var_\pi \left(n^{-1} \sum\nolimits_{i \in s} e_i / p_i \right) \\ &= n^{-1} \sum\nolimits_{i \in U} p_i \left(e_i / p_i - E_U \right)^2 \end{aligned} . \qquad (11)$$

Steps 3 and 4. We follow Spencer's approach and substitute model values in variance (11) to formulate a design-effect measure. However, here we substitute in the model-based equivalent to e_i, not y_i as Spencer does. 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, $Var_{srs} \left(\hat{T}_{pwr} \right) = N^2 \sigma_y^2 / n$, where $\sigma_y^2 = N^{-1} \sum\nolimits_{i=1}^{N} \left(y_i - \bar{Y} \right)^2$, will produce our approximate design effect due to unequal calibration weighting.

We can simplify things greatly 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 is

$$Deff_H = \frac{n\bar{W}}{N}\left(\frac{\sigma_u^2}{\sigma_y^2}\right) + \frac{n\sigma_w}{N\sigma_y^2}\left(\rho_{u^2w}\sigma_{u^2} - 2A\rho_{uw}\sigma_u\right). \tag{12}$$

where $\sigma_u^2 = N^{-1}\sum_{i=1}^{N}\left(u_i - \bar{U}\right)^2$, $\sigma_y^2 = N^{-1}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2$, ρ_{u^2w} 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.

Step 5. To estimate (12), we use

$$deff_H \approx deff_K\left(\mathbf{w}\right)\frac{\hat{\sigma}_u^2}{\hat{\sigma}_y^2} + \frac{n\hat{\sigma}_w}{N\hat{\sigma}_y^2}\left(\hat{\rho}_{u^2w}\hat{\sigma}_{u^2} - 2\hat{\alpha}\hat{\rho}_{uw}\hat{\sigma}_u\right), \tag{13}$$

where the model parameter estimate $\hat{\alpha}$ is obtained using survey-weighted least squares, $\hat{\sigma}_y^2$ and $\hat{\bar{y}}_w$ were defined in section 2.3, $\hat{\sigma}_u^2 = \sum_{i\in s} w_i\left(\hat{u}_i - \hat{\bar{u}}_w\right)^2 \Big/ \sum_{i\in s} w_i$, $\hat{\bar{u}}_w = \sum_{i\in s} w_i\hat{u}_i \Big/ \sum_{i\in s} w_i$, and $\hat{u}_i = y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$.

If the correlations in (12) are negligible, then we obtain $Deff_H \approx \frac{n\bar{W}}{N}\left(\frac{\sigma_u^2}{\sigma_y^2}\right)$, which can be estimated with

$$deff_H \approx deff_K\left(\mathbf{w}\right)\hat{\sigma}_u^2 \big/ \hat{\sigma}_y^2. \tag{14}$$

Note that without calibration, 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 (14) becomes $Deff_H \approx n\bar{W}/N$, which we estimate with Kish's measure $deff_K = 1 + \left[CV\left(\mathbf{w}\right)\right]^2$. 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 (14) is smaller than Kish's estimate using only the weights. 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 with both a one-sample example case study and simulation that mimics establishment-type data. We also examine the extent to which the correlation components in our proposed design effect (12) are significant, or large enough to influence the exact measure. Our design effect measure is a model-based version of the standard measure (4). Its calculation requires only the sample y-values, covariates, and calibration weights. This measure can, thus, be produced more quickly than measure (3), whose components are often available later in data processing after a variance estimation system is set up.

4. Evaluation Using Establishment Data

Here a sample dataset of tax return data is used to mimic an establishment survey setup. The data come from the Tax Year 2007 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 2010. 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" example 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 purposes. Four variables of interest are used: Total Assets, Total Liabilities, Total Revenue, and Total Expenses. Returns that were sampled with certainty and having "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 up to a pseudopopulation of 50,000 units.

Figure 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.

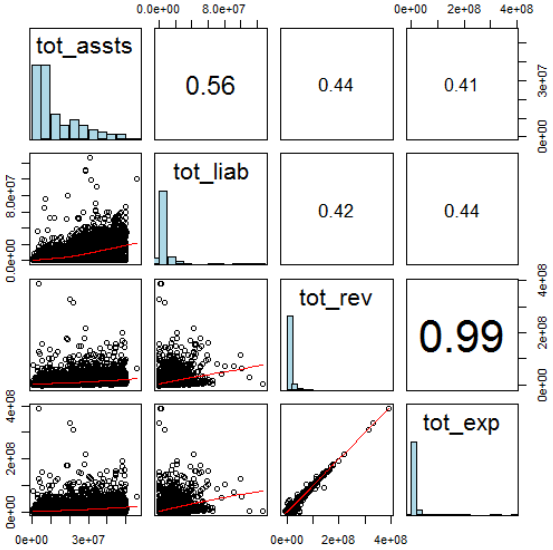

Figure 1. Pseudopopulation Values and Loess Lines for Design Effect Evaluation

4.1 One-Sample Example Results

Three *pps* 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. 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. Figures 2 and 3 show boxplots and plots of the sample weights before (labeled "HT wt" in Fig. 2) and after ("cal wt") these adjustments.

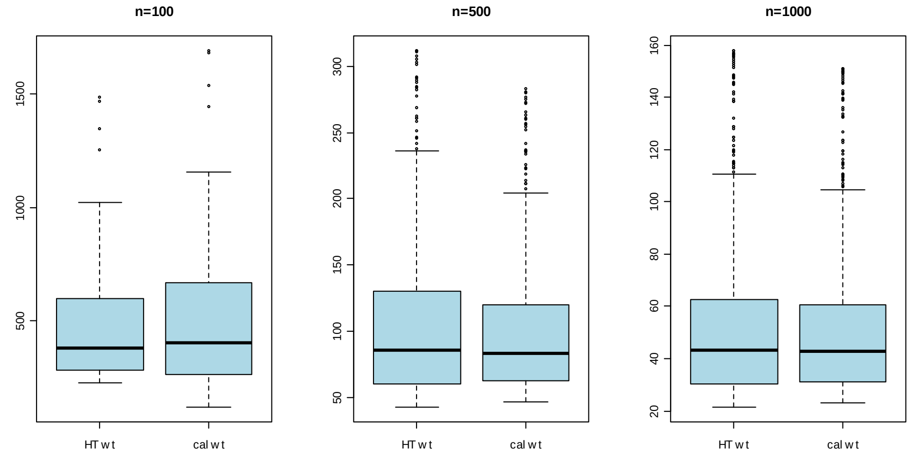

Figure 2. Boxplots of *PPSWR* Sample Weights Before and After Calibration Adjustments

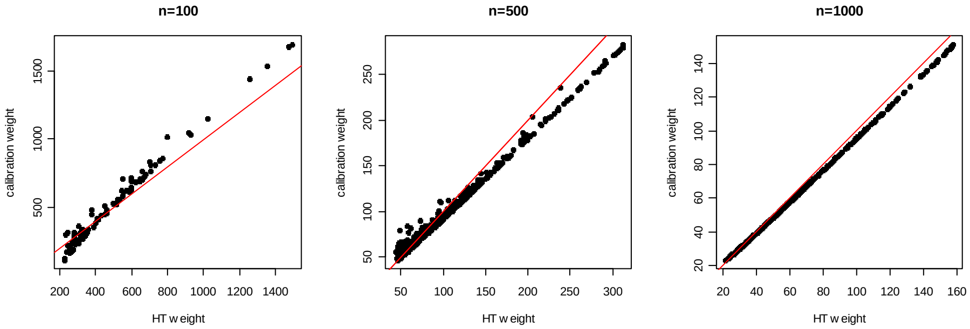

Figure 3. Plots of *PPSWR* Sample Weights Before and After Calibration Adjustments

As is apparent from Figure 3, the HT weights and the GREG weights do not differ dramatically; however, the GREG estimator is more efficient as shown below. Eight estimates of the design effects are considered, with results shown in Table 1:

- The standard design effect measures (2) and (3). Expression (2) reflects the efficiency of *pps* sampling and use of the π-estimator. Expression (3) reflects gains (if any) of *pps* sampling combined with GREG estimation;

- The Kish measure (4) computed using the GREG weights;

- Three Spencer measures: (i) the exact measure that estimates (5), (ii) the approximation (7) assuming zero correlation terms, and (iii) the large-population approximation (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 (13) and (ii) the zero-correlation approximation (14). 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 also use the conventional approach of selecting samples without replacement but computing estimated design effects that refer to with-replacement sampling. Several results are clear from Table 1. First, use of *PPSWR* sampling and the HT-estimator is more efficient than *srswr* in this population. $deff\left(\hat{T}_{HT}\right)$ ranges from 0.70 to 0.83, depending on the variable and sample size. The GREG is estimated to be considerably more efficient with *deff*'s of 0.34 to 0.39 for Liabilities and from 0.01 to 0.02 for Expenses. For these three *PPSWR* samples, the Kish measure is consistently above one for all sample sizes. This measure also does not depend on the variable of interest, and the fact that all estimates exceed one incorrectly implies that the *pps* sample design with or without calibration weighting is quite inefficient.

For Liabilities and Expenses, the Spencer exact measure ranges from 0.27 to 0.45 and overstates the efficiency of the (*PPSWR* sampling, *pwr*-estimator) combination compared to the directly computed *deff* of the HT estimator. However, the Spencer zero-correlation and large-N approximations are also inaccurate. For example, the zero-correlation and large-N approximations for (Liabilities, n=100) are 1.04 and 1.06 but $deff\left(\hat{T}_{HT}\right)$=0.83 for that variable and sample size. The exact proposed measure is exactly close, within two decimals rounding, to the directly calculated $deff\left(\hat{T}_{GREG}\right)$ in all cases.

Table 1. Design Effect Estimates of Single *PPSWR* Samples Drawn from the SOI 2007 Pseudopopulation EO Data

	Variable of Interest					
	Total Liabilities (weakly correlated with $\mathbf{X}$)			Total Expenses (strongly correlated with $\mathbf{X}$)		
Design Effect Estimates	$n=100$	$n=500$	$n=1000$	$n=100$	$n=500$	$n=1000$
Direct design effects						
HT-estimator[*]	0.83	0.75	0.70	0.72	0.75	0.78
GREG estimator[**]	0.34	0.34	0.39	0.02	0.02	0.01
Kish	1.42	1.25	1.26	1.42	1.25	1.26
Spencer						
Exact	0.27	0.45	0.45	0.51	0.63	0.53
Zero-corr. approx.	1.04	0.93	0.97	1.08	1.02	1.07
Large-*N* approx.	1.06	0.89	0.95	1.18	1.07	1.09
Proposed						
Exact	0.34	0.35	0.39	0.02	0.02	0.01
Zero-corr. approx.	0.85	0.66	0.79	0.05	0.03	0.02

[*] $Var_{\pi}\left(\hat{T}_{\pi}\right)/Var_{srs}\left(\hat{T}_{\pi}\right)$; [**] $Var_{\pi}\left(\hat{T}_{GREG}\right)/Var_{srs}\left(\hat{T}_{\pi}\right)$; both measures calculated with R's svytotal function.

We can understand why calibration is more efficient for Expenses than for Liabilities by examining the distributions of y_i and u_i. Figures 4 and 5 show boxplots of u_i and y_i for each variable and sample size. We see that, particularly for the Total Expenses variable, the u_i-values in all of these samples have shorter ranges of values and less variation than y_i. This occurs since Total Expenses is highly correlated with the calibration variable Total Revenue (see Figure 1) and explains why the direct and proposed design effect measures are so much smaller for Total Expenses.

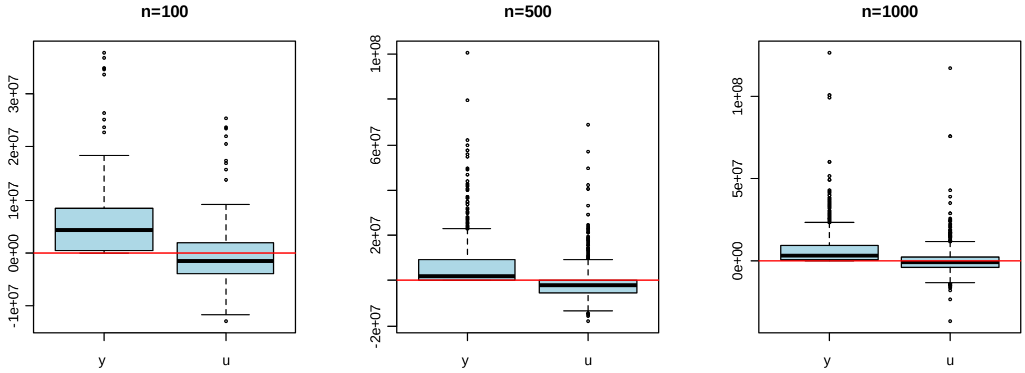

Figure 4. Boxplots of y_i and u_i -values from *ppswr* Samples from the 2007 SOI EO Data, Total Liabilities Variable (weakly correlated with $\mathbf{x}$)

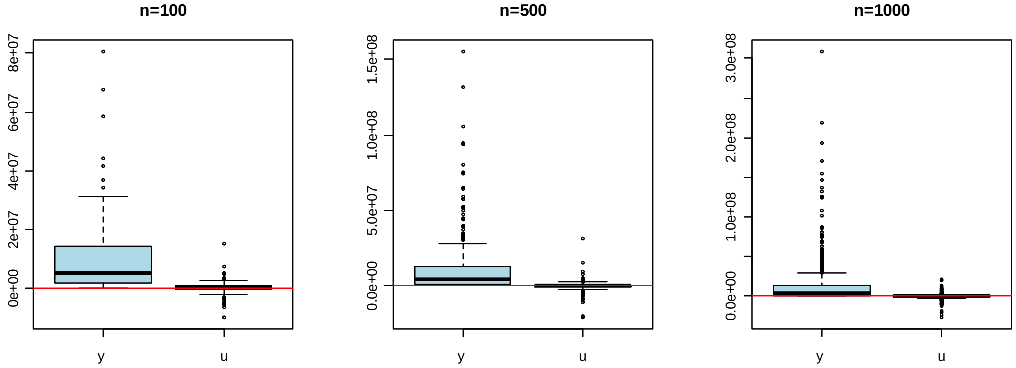

Figure 5. Boxplots of y_i and u_i -values from *ppswr* Samples from the 2007 SOI EO Data, Total Expenses Variable (strongly correlated with $\mathbf{x}$)

4.2. Simulation Study Results

We replicated the *PPSWR*-sampling in the previous section 10, 000 times 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

$$relbias\left(\hat{T}_s\right) = 100 \times \sum_{s=1}^{S} \left(\hat{T}_s - T\right) \Big/ T$$

$$MSE\ ratio = MSE\left(\hat{T}_{HT}\right) \Big/ MSE\left(\hat{T}_{GREG}\right)$$

$$= \sum_{s=1}^{S} \left(\hat{T}_{HT,s} - T\right)^2 \Big/ \sum_{s=1}^{S} \left(\hat{T}_{GREG,s} - T\right)^2$$

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 $\hat{T}_{HT,s}$ and $\hat{T}_{GREG,s}$ are the estimated HT and GREG totals from sample *s*. These results are shown in Table 2 on the following page.

Table 2. Simulation Results of HT and GREG Totals, 10,000 *ppswr* Samples Drawn from the SOI 2007 Pseudopopulation EO Data

	Variable of Interest					
	Total Liabilities (weakly correlated with $\mathbf{X}$)			Total Expenses (strongly correlated with $\mathbf{X}$)		
Estimates	$n=100$	$n=500$	$n=1000$	$n=100$	$n=500$	$n=1000$
relbias(HT)	0.04	0.06	0.03	0.06	0.03	-0.03
relbias(GREG)	0.47	0.13	0.13	-0.10	-0.01	0.00
MSE ratio	1.20	1.19	1.21	38.99	41.56	42.15

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=1000.

As seen in Table 2, both estimators 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. We also computed the biases of the various estimated design effects across the 10,000 samples. The relbiases of the Kish design effect estimates are computed as

$$relbias\left(deff_K\right)=100\times\left(\overline{deff_K}-empdeff\left(\hat{T}_{HTy}\right)\right)\Big/empdeff\left(\hat{T}_{HTy}\right),$$

where $\overline{deff_K}$ is the average Kish *deff* over all samples and $empdeff\left(\hat{T}_{HTy}\right)$ is the average over all samples of the *deff*'s of $\hat{T}_{HTy}$ computed from the `survey` package. The relbiases of the Spencer and proposed measures are computed as

$$relbias\left(deff_S\right)=100\times\left(\overline{deff_S}-empdeff\left(\hat{T}_{HTy}\right)\right)\Big/empdeff\left(\hat{T}_{HTy}\right) \text{ and}$$

$$relbias\left(deff_H\right)=100\left(\overline{deff_H}-empdeff\left(\hat{T}_{GREG}\right)\right)\Big/empdeff\left(\hat{T}_{GREG}\right)$$

where $\overline{deff_S}$ and $\overline{deff_H}$ are, respectively, the means of one of the Spencer or proposed alternatives. $empdeff\left(\hat{T}_{GREG}\right)$ is computed as the average over all samples of the *deff*'s of the GREG computed from the `survey` package. The relbiases are displayed in Table 3.

Table 3. Relative Bias of Design Effect Estimates,10,000 Samples Drawn from the SOI 2007 Pseudopopulation EO Data

	Variable of Interest					
	Total Liabilities (weakly correlated with $\mathbf{X}$)			Total Expenses (strongly correlated with $\mathbf{X}$)		
Design Effect Estimates	$n=100$	$n=500$	$n=1000$	$n=100$	$n=500$	$n=1000$
Empirical *deff*'s*						
HT	0.51	0.50	0.51	0.65	0.63	0.64
GREG	0.43	0.42	0.42	0.02	0.02	0.02
Relative biases (percent)	75.2	75.2	75.3	66.9	66.0	65.1
Kish**						
Spencer**	-46.2	-41.0	-39.4	-25.6	-25.8	-24.9
Exact	32.8	34.3	34.6	37.1	40.0	40.1
Zero-corr. approx.	33.2	34.9	35.2	44.7	46.6	46.5
Large-*N* approx.						
Proposed***	-0.2	0.2	0.2	11.5	-0.4	-0.1
Exact	94.0	96.8	97.6	104.1	91.0	93.0
Zero-corr. approx.						

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

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

Figures 6 and 7 at the end of this paper show boxplots of the alternative empirical design effect estimates. Each plot also shows vertical reference lines for the average of the empirical HT (in red) and GREG (in blue) design effects. We can see that the empirical distributions are all skewed and the proposed exact design effect cover the average of the GREG design effect for both variables, particularly Total Expenses.

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 a calibration estimator.

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, so we recommend incorporating them when possible to avoid overly high estimates of the design effects. However, the high correlations between survey and auxiliary variables that we observed in our 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 (11) is approximately correct in large samples for all calibration estimators (Särndal *et. al* 1992), 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.

References

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

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.

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.

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 D. Pfeffermann and C. R. Rao (Eds.), *Handbook of Statistics*, *Sample Surveys: Design, Methods and Application,* **29B**, Amsterdam: Elsevier BV.

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

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

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

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

Spencer, B. D. (2000), “An Approximate Design Effect for Unequal Weighting When Measurements May Correlate With Selection Probabilities,” *Survey Methodology,* **26**, 137-138.

Statistics of Income (2011), “2007 Charities & Tax-Exempt Microdata Files,” available at: http://www.irs.gov/taxstats/article/0,,id=226223,00.html.

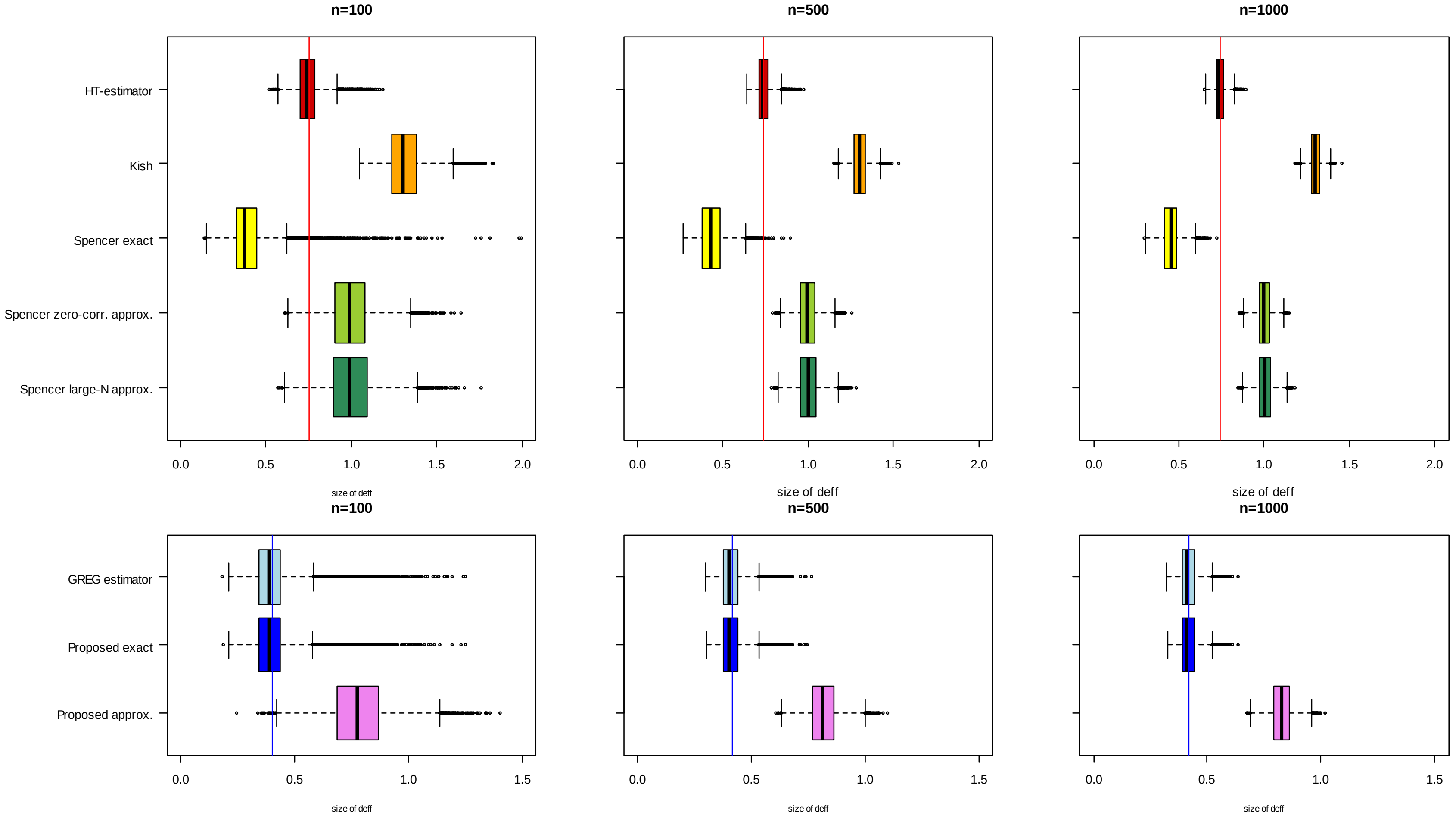

Figure 6. Empirical Boxplots of Design Effect Estimates of 10,000 *PPSWR* Samples Drawn from the SOI 2007 Pseudopopulation EO Data Total Liabilities Variable (weakly correlated with $\mathbf{x}$). Red line is average of empirical HT *deff*'s; blue line average of empirical GREG *deff*'s.

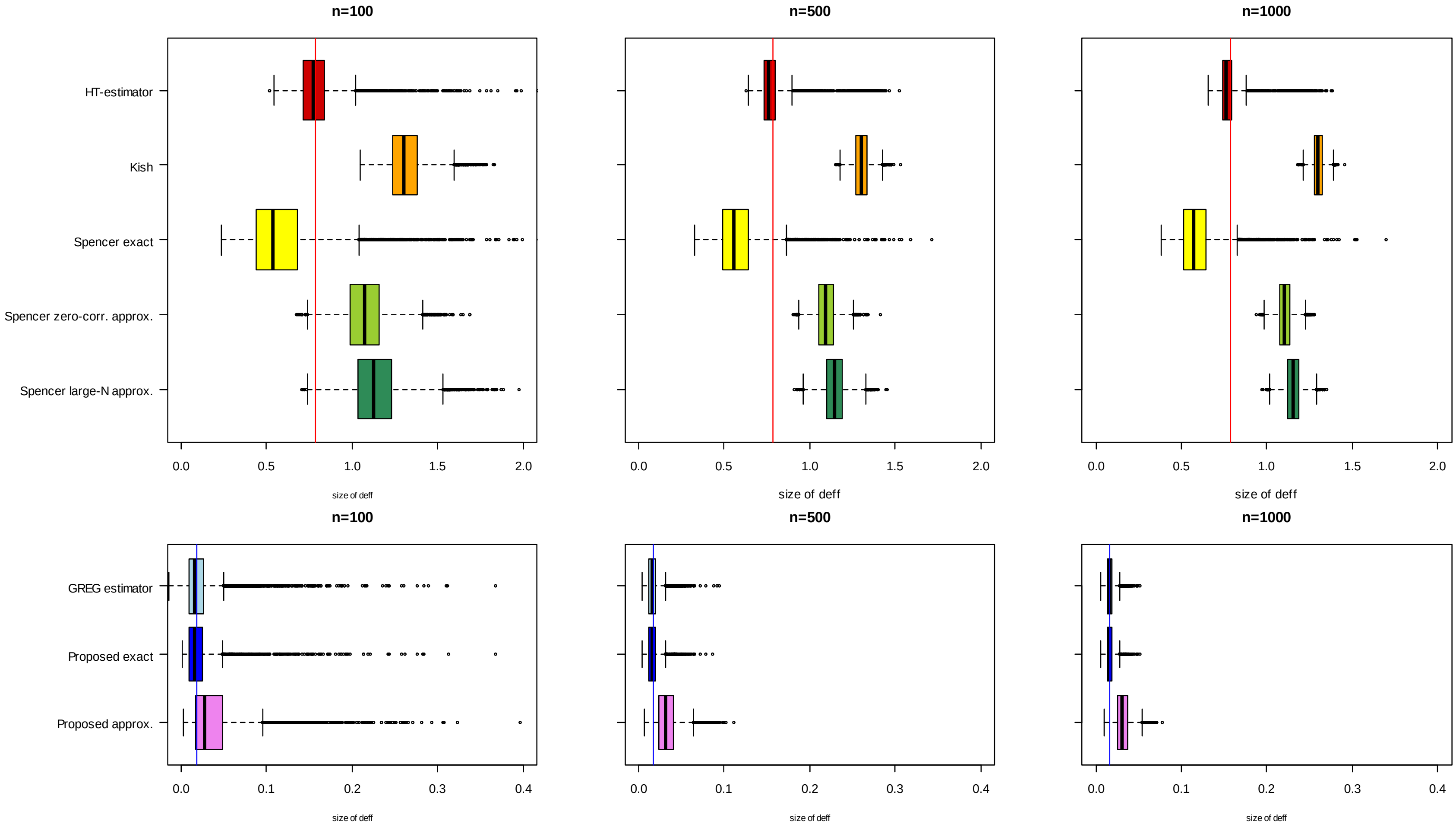

Figure 7. Empirical Boxplots of Design Effect Estimates of 10,000 *PPSWR* Samples Drawn from the SOI 2007 Pseudopopulation EO Data Total Expenses Variable (strongly correlated with $\mathbf{x}$). Red line is average of empirical HT *deff*'s; blue line average of empirical GREG *deff*'s.