A Design Effect Measure for Calibration Weighting in Cluster 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 for two-stage sampling when calibration weighting is used. 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. The proposed measure is compared to existing design effect measures in an example involving household-type data.

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

1. Introduction

The most popular measure for gauging the effect of differential weighting on the precision of an estimator is Kish's (1965, 1992) design-based design effect. Gabler *et al.* (1999) showed that, for cluster sampling, this estimator is a special form of a design effect produced using variances from random effects models, with and without intra-class correlations. Spencer (2000) proposed a simple model-based approach that depends on a single covariate to estimate the impact on variance of using variable weights.

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 the Kish design effects attempt to measure the impact of variable weights, as noted in Kish (1992), 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. 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. Henry and Valliant (2013) extended this to gauge the impact of differential calibration weights in single-stage samples.

In particular, the existing 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 element-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 two-stage cluster-sample 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, but also uses a generalized regression variance to reflect multiple calibration covariates and the cluster sample design. In section 4, we apply the estimators in a case study involving household-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 existing design effect measures for cluster sampling. The assumptions used to derive each of these are also presented.

2.1. Notation

We consider that a finite population of M elements is partitioned into N clusters, with sizes M_i, and denoted by $U=\left\{(i,j): i=1,\ldots,N,\ j=1,\ldots,M_i\right\}$. On element (i,j), an analysis variable y_{ij} is observed. The population total of the y's is $T=\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}$ and the population variance is $\sigma_y^2=M^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M_i}\left(y_{ij}-\bar{Y}\right)$, where $M=\sum_{i=1}^{N}M_i$ and $\bar{Y}=T/M$. We select a sample of n clusters and m_i elements within cluster i, using two-stage sampling from U and obtain a set of $s=\left\{(i,j): i=1,\ldots,n,\ j=1,\ldots,m_i\right\}$ respondents. The total sample size of elements is $m=\sum_{i=1}^{n}m_i$.

In the following sections, when a two-stage design is considered, we assume that clusters are selected with replacement. The selection probability of cluster i on a single draw is p_i. Within cluster i, the sample of m_i elements is selected via simple random sampling without replacement (*srswor*). Assuming that we have probability-with-replacement (*pwr*) sampling of clusters, the probability of selection for clusters is approximately $\pi_i=1-\left(1-p_i\right)^n \doteq np_i$ (if p_i is not too large), where p_i is the one-draw selection probability. The second-stage selection probability is $\pi_{j|i}=m_i/M_i$ for element j in cluster i. Then the overall selection probability is approximately $\pi_{ij}=\pi_i\pi_{j|i}\doteq np_i m_i/M_i$. We estimate the total from the sample using with the *pwr*-estimator $\hat{T}_{pwr,y}=\sum_{i=1}^{n}\frac{1}{np_i}\sum_{j=1}^{m_i}\frac{M_i}{m_i}y_{ij}\equiv\sum_{i=1}^{n}\sum_{j=1}^{m_i}w_{ij}y_{ij}$, where $w_{ij}=M_i/\left(np_i m_i\right)$.

2.2. 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 *m* elements is selected with π_i being the selection probability of element *i* and $\mathbf{x}_i$ a vector of *p* auxiliaries associated with element *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 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 element 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 (Deville and Särndal 1992) when the sample size is large. This result holds for multistage sampling and allows us to produce one design effect measure applicable to all estimators in the family of calibration estimators.

2.3. Direct Design-Effect Measures

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

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

where $\hat{T}_{srswr} = M \sum\nolimits_{k \in s} y_k \Big/ m$ is the expansion estimator under simple random sampling with replacement (*srswr*) and $Var\left(\hat{T}_{srswr}\right) = M^2 \sigma_y^2 \Big/ m$. 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 an *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 (Deville and Särndal 1992), in which case

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

As described in Sec. 3, our proposed design effect is a model-based approximation to (3).

2.4. Kish's "Haphazard-Sampling" Design-Effect for Unequal Single-stage Sample 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 $\mathbf{w} = \left(w_1, \ldots, w_m\right)^T$ denoting the weights from a simple random sample without replacement (*srswor*) of m elements, this measure is

$$\begin{aligned} deff_K\left(\mathbf{w}\right) &= 1 + \left[CV\left(\mathbf{w}\right)\right]^2 \\ &= m\sum\nolimits_{i\in s} w_i^2 \Big/ \left[\sum\nolimits_{i\in s} w_i\right]^2, \end{aligned} \tag{4}$$

where $CV\left(\mathbf{w}\right) = \sqrt{m^{-1}\sum\nolimits_{i\in s}\left(w_i - \bar{w}\right)^2 \Big/ \bar{w}^2}$ is the coefficient of variation of the weights, and $\bar{w} = m^{-1}\sum\nolimits_{i\in s} w_i$. Expression (4) is derived from the ratio of the variance of the weighted survey mean under disproportionate stratified *srswor* to the variance under proportionate stratified *srswor* when all stratum element variances are equal (Kish 1992). With equal stratum variances, sampling with a proportional allocation to strata is optimal, which leads to all elements having the same weight.

2.5. Kish's Measure for Cluster Sampling Weights

Kish (1987) proposed a similar measure for cluster sampling. Assume that there are G unique weights in s such that the m_{ig} elements within each cluster i have the same weight, denoted by $w_{ig} = w_g$ for $g = 1,\ldots,G$, m_g is the number of elements within weighting class g and $m = \sum\nolimits_{g=1}^{G} m_g$ is the total number of elements in the sample. We estimate the population mean $\bar{Y} = T/M$ using the weighted sample mean $\bar{y}_w = \sum\nolimits_{g=1}^{G}\sum\nolimits_{j\in g} w_{gj} y_{gj} \Big/ \sum\nolimits_{g=1}^{G}\sum\nolimits_{j\in g} w_{gj}$. Kish's (1987) decomposition model for $\bar{y}_w$ assumes that the G weighting classes are randomly ("haphazardly") formed with respect to y_{ij}, assuming that the y_{ij} have a common variance and that s is an equal-probability sample in which the variation among the m_g's within s is not significant. The resulting design effect is

$$deff_{KC}\left(\bar{y}_w\right)=\frac{m\sum_{g=1}^{G}w_g^2 m_g}{\left[\sum_{g=1}^{G}w_g m_g\right]^2}\times\left[1+\left(\bar{m}-1\right)\rho_c\right], \tag{5}$$

where $\bar{m}=n^{-1}\sum_{i=1}^{n}m_i$ is the average cluster size and ρ_c the measure of intra-cluster homogeneity. The first component in (5) is the cluster-sample equivalent of (4), and can be written in a similar form using the squared CV of the weights if $m_g=1$. The second (5) component is the standard design effect due to cluster sampling (e.g., Kish 1965). Expression (5) may not hold if there is variation in the m_{ig} across clusters (Park 2004) or moderate correlation between the survey characteristic and weights (Park and Lee 2004).

2.6. Gabler et al.'s Measure for Cluster Sampling

Gabler *et al*. (1999) used a model to justify measure (5) that assumes y_{ij} is a realization from a one-way random effects model (i.e., a one-way ANOVA-type model with only a random cluster-level intercept term plus an error) that assumes the following covariance structure:

$$Cov\left(y_{ij},y_{i'j'}\right)=\begin{cases}\sigma^2 & i=i', j=j'\\ \rho_e\sigma^2 & i=i', j\neq j'\\ 0 & i\neq i'\end{cases}. \tag{6}$$

If the elements are uncorrelated, then (6) reduces to $Cov\left(y_{ij},y_{i'j'}\right)=\sigma^2$ for $i=i', j=j'$ and 0 otherwise. Gabler *et al.* (1999) take the ratio of the model-based variance of the weighted survey mean under a model with covariance structure (6) to the variance under the uncorrelated errors version and derive

$$deff\left(\bar{y}_w\right)=\frac{m\sum_{g=1}^{G}w_g^2 m_g}{\left[\sum_{g=1}^{G}w_g m_g\right]^2}\times\left[1+\left(\bar{m}_{g_1}-1\right)\rho_e\right] \quad , \tag{7}$$

where $\bar{m}_{g_1}=\sum_{i=1}^{n}\left(\sum_{g=1}^{G}w_g m_{ig}\right)^2\Big/\sum_{g=1}^{G}w_g^2 m_g$. They also established an upper bound for (7):

$$UB\left[deff\left(\bar{y}_w\right)\right]=\frac{m\sum_{g=1}^{G}w_g^2 m_g}{\left[\sum_{g=1}^{G}w_g m_g\right]^2}\times\left[1+\left(\bar{m}_{g_2}-1\right)\rho_e\right] \quad , \tag{8}$$

where $\bar{m}_{g_2}=\sum_{i=1}^{n}m_i\sum_{g=1}^{G}w_g^2 m_{ig}\Big/\sum_{i=1}^{n}\sum_{g=1}^{G}w_g^2 m_{ig}$ is a weighted average of cluster sizes.

Park (2004) further extends this approach to three-stage sampling, assuming that a systematic sampling is used in the first stage to select the clusters. Lynn and Gabler (2005) provide examples of special cases of (7), such as equal sampling/coverage/response rates across domains.

Lynn and Gabler (2005) rewrite $\bar{m}_{g_2} = N \dfrac{Cov\left(m_g, m_g \bar{w}_g^2\right) + \bar{m}\sum_{g=1}^{G} m_g \bar{w}_g^2}{\sum_{g=1}^{G} m_g Var\left(w_{ig}\right) + \sum_{g=1}^{G} m_g \bar{w}_g^2}$ in Gabler *et al*. (1999)'s design effect. Assuming certain restrictions on the weights, they propose a survey planning approximation, $\bar{m}_{g_3} = m\left(1+\left[CV\left(\mathbf{m}\right)\right]^2\right) \Big/ n\left(1+\left[CV\left(\mathbf{w}\right)\right]^2\right)$, where $1+\left[CV\left(\mathbf{m}\right)\right]^2 = G\sum_{g=1}^{G} m_g^2 \Big/ \left(\sum_{g=1}^{G} m_g\right)^2$ uses the squared CV of cluster sample sizes across clusters, and $1+\left[CV\left(\mathbf{w}\right)\right]^2 = \sum_{g=1}^{G} w_g^2 m_g \Big/ \left(\sum_{g=1}^{G} w_g m_g\right)^2$ uses the squared CV of weights across observations.

2.7. Park and Lee's Measures for Unequal Cluster Sampling

Park and Lee (2004) extend the Gabler *et al*. (1999) design effect to account for unequal sampling weights within a two-stage cluster sample that selects $\bar{m} \geq 2$ elements from each PSU:

$$deff\left(\hat{T}_{HTy}\right) = 1 + \left(\bar{m}-1\right)\tau + W_y^*, \tag{9}$$

where $\tau = \dfrac{\left(N-1\right)S_{yB}^2 + \sum_{i=1}^{N}\left(\bar{m}-1\right)^{-1} S_{yi}^2}{\left(N-1\right)S_{yB}^2 + \sum_{i=1}^{N}\left(M_i-1\right)S_{yi}^2}$, $S_{yi}^2 = \left(M_i - 1\right)^{-1}\sum_{j=1}^{M_i}\left(y_{ij} - \bar{Y}_i\right)^2$ is the within-cluster variability, $S_{yB}^2 = \left(N-1\right)^{-1}\sum_{i=1}^{N} M_i\left(\bar{Y}_i - \bar{Y}\right)^2$ is the across-cluster variability among cluster means, $W_y^* = \dfrac{m_0}{CV_y^2}\sum_{i=1}^{N}\left(\dfrac{Q_i}{p_i M^2}\right)\dfrac{\bar{Y}_i^2}{\bar{Y}^2}\left(1+\dfrac{CV_{yi}^2}{m_0}\right)$, $CV_y^2 = \sigma_y^2 \big/ \bar{Y}^2$, $CV_{yi}^2 = S_{yi}^2 \big/ \bar{Y}_i^2$, and $\bar{Y}_i = \sum_{i=1}^{M_i} y_{ij} \big/ M_i$. The term $Q_i = M_i\left(M_i - p_i M\right)$ is equal to zero if $p_i = M_i / M$ or the cluster probability of selection is proportional to the cluster size M_i. The equivalent of (9) for the weighted sample mean is

$$deff_{PL}\left(\bar{y}_w\right) = 1 + \left(m_0 - 1\right)\tau + W_d^*, \tag{10}$$

where $W_d^* = m_0 \Big/ CV_y^2 \sum_{i=1}^{N}\left(Q_i \big/ p_i M^2\right)\left(\bar{D}_i^2 \big/ \bar{Y}^2\right)\left(1 + CV_{di}^2 \big/ m_0\right)$, $CV_{di}^2 = S_{di}^2 \big/ \bar{D}_i^2$, $\bar{D} = \sum_{i=1}^{N}\sum_{i=1}^{M_i} d_{ij} \Big/ M$, and $\bar{D}_i = \sum_{j=1}^{M_i} d_{ij} \big/ M_i$ for the transformed variable $d_{ij} = y_{ij} - \bar{Y}$.

2.8. Spencer's Model-based Measure for Single-Stage *pwr* 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. This measure was not developed for cluster sampling, but we used this modeling approach in Sec. 3 for a proposed design effect. The sample is assumed to be selected with varying probabilities and with replacement (*pwr*). Suppose that p_i is the one-draw probability of selecting element 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 element 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 *pwr*-sampling is $w_i = \left(np_i\right)^{-1}$. The estimated total studied by Spencer is $\hat{T}_{pwr,y} = \sum_{i \in s} w_i y_i$, with design-variance $Var\left(\hat{T}_{pwr,y}\right) = n^{-1} \sum_{i \in U} p_i \left(y_i / p_i - T\right)^2$ 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; modified for our notation):

$$Deff_S = \frac{A^2}{\sigma_y^2}\left(\frac{m\bar{W}}{M} - 1\right) + \frac{m\bar{W}}{M}\left(1 - \rho_{yp}^2\right) + \frac{m\rho_{e^2 w}\sigma_{e^2}\sigma_w}{M\sigma_y^2} + \frac{2Am\rho_{ew}\sigma_e\sigma_w}{M\sigma_y^2}. \tag{11}$$

Assuming that the correlations in the last two terms of (11) are negligible, Spencer approximates (11) with the first two terms in (11), 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. When ρ_{yp} is zero and σ_y is large, measure (11) 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.

3. Proposed Design Effect Measures

Henry and Valliant's (2013) approach in single-stage sampling can be extended to produce a new weighting design effect measure for a calibration estimator in cluster sampling. 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*.

First, a linearization approximation to the GREG estimator (Exp. 6.6.9 in Särndal *et al*. 1992) assuming clusters are selected *pwr* is

$$\begin{aligned} \hat{T}_{GREG} &= \hat{T}_{pwr,y} + \left(\mathbf{T}_x - \hat{\mathbf{T}}_{pwr,x}\right)^T \hat{\mathbf{B}} \\ &\doteq \hat{T}_{pwr,y} + \left(\mathbf{T}_x - \hat{\mathbf{T}}_{pwr,x}\right)^T \mathbf{B}_U \\ &= \sum_{i \in s} \sum_{j \in s_i} w_{ij} e_{ij} + \mathbf{T}_x^T \mathbf{B}_U \end{aligned} \tag{12}$$

where $\mathbf{T}_x$ is the known population total of $\mathbf{x}$, $\hat{\mathbf{T}}_{pwr,x}$ is the vector of *pwr*-estimators, $\mathbf{B}_U$ is the population coefficient, $e_{ij} = y_{ij} - \mathbf{x}_{ij}^T \mathbf{B}_u$ is the residual for element i, j, and $\mathbf{x}_{ij}^T$ is a row vector of calibration covariates. From (12), $\hat{T}_{GREG} - \mathbf{T}_x^T \mathbf{B}_U \doteq \sum_{i \in s} \sum_{j \in s_i} w_{ij} e_{ij}$. Under the two-stage sampling assumptions (Sec. 2.1), $w_{ij} = M_i \big/ \left(np_i m_i \right)$ and the approximation variance of the GREG in (12) is also the variance of

$$\begin{aligned} \hat{T}_{GREG} - \mathbf{T}_x^T \mathbf{B}_U &\doteq \sum_{i \in s} \sum_{j \in s_i} M_i e_{ij} \big/ np_i m_i \\ &= \sum_{i \in s} w_i \hat{T}_{ei} \end{aligned} , \qquad (13)$$

where $w_i = \left(np_i \right)^{-1}$ and $\hat{T}_{ei} = M_i / m_i \sum_{j \in s_i} e_{ij}$.

The design-variance of (13) is

$$\begin{aligned} Var\left(\hat{T}_{GREG} \right) &\doteq \frac{1}{n} \sum_{i=1}^{N} p_i \left(\frac{e_{Ui+}}{p_i} - E_{U+} \right)^2 + \sum_{i=1}^{N} \frac{M_i^2}{np_i m_i} \left(1 - \frac{m_i}{M_i} \right) \frac{1}{M_i - 1} \sum_{j=1}^{M_i} \left(e_{ij} - \bar{e}_{U_i} \right)^2 \\ &= \frac{1}{n} \sum_{i=1}^{N} \left(\frac{e_{Ui+}^2}{p_i} \right) - E_{U+}^2 + \sum_{i=1}^{N} \frac{M_i^2}{np_i m_i} \left(1 - \frac{m_i}{M_i} \right) S_{Uei}^2 \end{aligned} , \qquad (14)$$

where $E_{U+} = \sum_{i=1}^{N} e_{ij}$, $e_{Ui+} = \sum_{j \in U_i} e_{ij}$, $S_{Uei}^2 = M_i^{-1} \sum_{j=1}^{M_i} \left(e_{ij} - \bar{e}_{Ui} \right)^2$, and $\bar{e}_{Ui} = \sum_{j=1}^{M_i} e_{ij} \Big/ M_i$. The true *deff* for $\hat{T}_{GREG}$ is defined as

$$deff_{True}\left(\hat{T}_{GREG} \right) = Var\left(\hat{T}_{GREG} \right) \Big/ Var\left(\hat{T}_{srswr} \right). \qquad (15)$$

3.1. Valliant et al.'s Relative Variance Deff

Valliant, Dever, and Kreuter (2013, Sec. 9.2.3) give a design effect for totals in cluster samples using the ratio of the relative variance of the estimator of the total when clusters are selected *pwr* and $\bar{m}$ elements within clusters are selected via *srswor* over the equivalent expression under *srswr*. Their measure can be modified by using (14) which uses the GREG-based residuals. Assuming that $\bar{m}$ elements are selected in each cluster, the relative variance of the GREG (i.e., the variance of the GREG over the squared total) can be rewritten as

$$RelVar\left(\hat{T}_{GREG} \right) \doteq RelVar\left(\hat{T}_{srswr} \right) \times k_e \left[1 + \delta_e \left(\bar{m} - 1 \right) \right], \qquad (16)$$

where $RelVar\left(\hat{T}_{srswr} \right) = \sigma_y^2 \Big/ \left(n\bar{m}\bar{Y}^2 \right)$, $\delta_e = B_e^2 \Big/ \left(B_e^2 + W_e^2 \right)$, $B_e^2 = \sum_{i=1}^{N} p_i^2 \left(e_{Ui+} / p_i - E_{U+} \right)^2 \Big/ T^2$ is the between-cluster relative variance component, $W_e^2 = \sum_{i=1}^{N} M_i S_{Uei}^2 \Big/ T^2$ is the within-cluster component, $k_e = \left(B_e^2 + W_e^2 \right) \Big/ RelVar\left(T \right) = \left(B_e^2 + W_e^2 \right) \Big/ \left(\sigma_y^2 \big/ T^2 \right)$, where $e_{ij} = y_{ij} - \mathbf{x}_{ij}^T \mathbf{B}_U$ is the residual for element *ij*, $\mathbf{x}_{ij}^{T}$ is a row vector of calibration covariates (including an intercept if one is used), and $E_U = \sum_{i \in U} \sum_{j \in U_i} e_{ij}$. Using the form of the relative variance in (16), the associated design effect is

$$deff_{VDK}\left(\hat{T}_{cal}\right) = k_e\left[1 + \delta_e\left(\bar{m} - 1\right)\right]. \tag{17}$$

When calibration is more efficient for a given y-variable, the term k_e will be less than 1 and acts as a dampening factor, reducing the overall design effect.

3.2. An Alternative Deff

We follow Spencer's approach and substitute model values in variance (14) to formulate a design-effect measure. However, here we substitute in the model-based equivalent to e_i, not y_i, as Spencer does.

Suppose that the second-stage sampling fraction is negligible, i.e., $m_i / M_i \approx 0$. To match the theoretical variance formulation in (14), consider the model $y_{ij} = A_U + \dot{\mathbf{x}}_{ij}^{T}\mathbf{B}_U + e_{ij}$, where $A_U, \mathbf{B}_U$ are the finite population model parameters, and $\mathbf{x}_{ij}$ from previous sections is equal to $\mathbf{x}_{ij} = \left(1 \quad \dot{\mathbf{x}}_{ij}\right)$.

Similar to Henry and Valliant (2013), we simplify things by reformulating the model as $u_{ij} = y_{ij} - \dot{\mathbf{x}}_{ij}^{T}\mathbf{B}_U$, such that $e_{ij} = u_{ij} - \alpha_U$ and incorporate $\alpha_i = \alpha_U M_i$ not as a cluster-level (random) intercept, simply an algebraic expression. The model assuming a random intercept α_i and error term is equivalent to Gabler *et al.*'s (1999) random effects model.

Substituting the GREG residuals e_{ij} into the variance (14) and taking its ratio to the variance of the *pwr*-estimator in *srswr*, $Var_{srs}\left(\hat{T}_{srswr}\right) = M^2\sigma_y^2 / n\bar{m}$ will produce our approximate design effect due to unequal calibration weighting. Assuming that the M_i are large enough such that $M_i / \left(M_i - 1\right) \approx 1$ and the within-cluster sampling fractions are negligible, we obtain the following approximate design effect:

$$deff_{G1} \approx \frac{n\bar{m}\bar{W}N\left(\sigma_{u_+}^2 + \bar{U}^2 + \sigma_\alpha^2 + \bar{\alpha}^2 - 2\overline{U\alpha}\right)}{M^2\sigma_y^2} - \frac{n\bar{m}N^2\left(\bar{U} - \bar{\alpha}\right)^2}{nM^2\sigma_y^2}$$
$$+ \frac{n\bar{m}N\sigma_w\left(\rho_{u_+^2 w}\sigma_{u_+^2} + \rho_{\alpha^2 w}\sigma_{\alpha^2} - 2\rho_{u_+\alpha,w}\sigma_{u_+\alpha}\right)}{M^2\sigma_y^2} + \frac{n\bar{m}}{M^2\sigma_y^2}\sum\nolimits_{i=1}^{N}\frac{M_i^2 w_i S_{U_{ui}}^2}{m_i}. \tag{18}$$

where $u_{i+} = \sum_{j \in U_i}\left(y_{ij} - \dot{\mathbf{x}}_{ij}^{T}\mathbf{B}_U\right)$, $\bar{U} = N^{-1}\sum_{i=1}^{N} u_{i+}$, $\bar{\alpha} = N^{-1}\sum_{i=1}^{N}\alpha_i = \alpha_U\bar{M}$, $\bar{M} = N^{-1}\sum_{i=1}^{N} M_i$, $\overline{U\alpha} = N^{-1}\sum_{i=1}^{N} u_{i+}\alpha_i$, $S_{U_{ui}}^2 = \left(M_i - 1\right)^{-1}\sum_{j=1}^{M_i}\left(u_{ij} - \bar{u}_i\right)^2$,

$\sigma_{u_+}^2 = N^{-1}\sum_{i=1}^{N}\left(u_{i+} - \bar{U}\right)^2$, $\sigma_\alpha^2 = N^{-1}\sum_{i=1}^{N}\left(\alpha_i - \bar{\alpha}\right)^2$, $\sigma_w = N^{-1}\sum_{i=1}^{N}\left(W_i - \bar{W}\right)$, and $\sigma_{u_+\alpha} = N^{-1}\sum_{i=1}^{N}\left(u_{i+} - \bar{U}\right)\left(\alpha_i - \bar{\alpha}\right)$. The ρ-components are the finite population correlations between the terms within each subscript.

Some approximations to (18) exist. If the correlations in (18) are negligible, then we obtain $deff_{G0} \approx \frac{m\bar{W}}{M}\frac{N}{M}\left(\frac{\sigma_{u_+}^2 + \bar{U}^2 + \sigma_\alpha^2 + \bar{\alpha}^2 - 2\overline{U\alpha}}{\sigma_y^2}\right) - \frac{mN^2\left(\bar{U} - \bar{\alpha}\right)^2}{nM^2\sigma_y^2} + \frac{m}{M^2\sigma_y^2}\sum_{i=1}^{N}\frac{M_i^2 w_i S_{U_{ui}}^2}{m_i}$, which can be estimated with the correlation terms removed from (18). Assuming that M_i are close enough such that $M_i \approx \bar{M}$, $\alpha_i = M_i\alpha \doteq \bar{M}\alpha$, and $\sigma_\alpha^2 = 0$, then (18) becomes

$$deff_{G1} \approx \frac{N\bar{m}\bar{W}}{\bar{M}^2\sigma_y^2}\left(\frac{\sigma_{u_+}^2 + \left(\bar{U} - \bar{M}\alpha_U\right)^2}{\sigma_y^2}\right) - \frac{\bar{m}N^2\left(\bar{U} - \bar{\alpha}\right)^2}{\bar{M}^2\sigma_y^2} + \\ + \frac{n\bar{m}N\sigma_w\left(\rho_{u_+^2 w}\sigma_{u_+^2} - 2\bar{M}\alpha_U\sigma_{u+}\rho_{u_+ w}\right)}{\bar{M}^2\sigma_y^2} + \frac{n\bar{m}}{\bar{M}^2\sigma_y^2}\sum_{i=1}^{M}\frac{w_i S_{U_{ui}}^2}{m_i}. \tag{19}$$

The Kish measure is also a special case of (19), when there are no cluster-level effects. In particular, suppose that for all elements $\dot{\mathbf{x}}_{ij} = 0$, and there is no cluster sampling. Then, $u_{ij} = y_{ij}$, $\sigma_{u+}^2 \approx \sigma_y^2$, $\bar{U} = \bar{M}\alpha = \bar{M}\bar{Y}$, $\sigma_{u_+}^2 = \sigma_y^2$, and $\sigma_{\alpha^2} = \sigma_\alpha^2 = 0$, and (19) reduces to $n\bar{m}\bar{W}/M$. The estimator of this from a particular sample is Kish's measure, $deff_K\left(\mathbf{w}\right)$, defined in (4).

When the relationship between the calibration covariates $\mathbf{x}$ and y is strong, the variance σ_{u+}^2 should be smaller than σ_y^2. In this case, measure (18) is smaller than the Kish and Park and Lee estimates. Variable weights produced from calibration adjustments are thus not as "penalized" (shown by overly high design effects) as they would be using the Kish. 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. This is illustrated in Sec. 4 with a population that mimics household-type data. We also examine the extent to which the correlation components in our proposed design effect (18) are significant, or large enough to influence the exact measure.

4. Empirical Example Using Household Data

The `MDarea.pop` dataset in the R `PracTools` package (Valliant *et al*. 2014) is used in this section as an example. This dataset contains 403,997 persons generated from data from the 2000 decennial U.S. Census for Anne Arundel County in the state of Maryland and is described in detail in Valliant *et al.* (2013). Individual values for each person were generated using models. Groupings to form the variables `PSU` and `SSU` were done after sorting the census file by tract and block group within tract.

We treated block groups within Census tracts as the cluster (PSU) and selected persons as the elements (SSU). PSUs with a small number of persons (less than 500) were excluded, leaving a pseudo-population of 274 PSU's and 397,065 elements. A poststratified estimator was used, with poststrata defined by the cross-classification of gender and 15 age groups (less than 5, 5-9, 10-14, 15-17, 18-21, 22-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65 and older). Thus, the weighting groups are post-strata, which cut across the PSU's and SSU's.

We examine six variables of interest, `y1`, `y2`, and `y3,` fictitious continuous variables on the file, created an additional variable `y4` using a linear model with the poststratification variables as covariates, and two binary variables indicating presence/absence of insurance coverage and a hospital stay. 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 household-type data patterns.

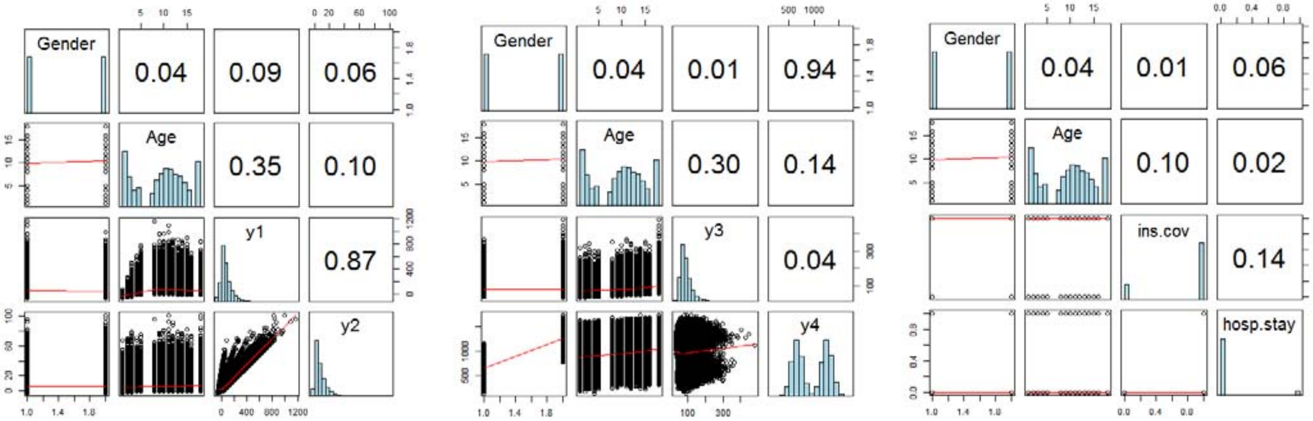

Figure 1. Pseudopopulation Values and Loess Lines for Design Effect Evaluation

Eight design effects are compared, with results shown in Table 1:

- The direct design effect measure $deff_{True}$ computed from (15). This reflects the combined effects of cluster sampling and poststratification;
- The Kish measure $deff_K$ (4) computed using the GREG weights
- The Kish measure $deff_{KC}$ (5) for cluster sampling;
- Park and Lee's measure $deff_{PL}$ in (10);
- $deff_{VDK}$ in (17);

- Three proposed measures: (i) $deff_{G1}$, the exact proposed design effect in (18), (ii) $deff_{G0}$, the zero-correlation approximation to $deff_{G1}$, and $deff_{G2}$, the equal-cluster sized approximation (19). All of these are meant to show the precision gains (if any) of the cluster sampling combined with GREG estimation.

Note that we do not include the Gabler *et al.* (1999) or Lynn and Gabler (2005) measures, which correspond to design effects of the weighted survey mean. Park and Lee (2004) describe the circumstances in which design effects for means and totals are different. Since our proposed design effect is for estimation of totals, we focus our empirical comparison on comparing it to only the existing measures for totals.

The results are shown in Table 1.

Table 1. Population Design Effects, by Variable of Interest

	Variable of Interest					
Design Effects	y_1*	y_2*	y_3*	y_4*	Hospital Stay**	Insurance Coverage**
$deff_{True}$	0.9	1.4	5.2	0.1	1.3	1.1
Kish						
Single-stage $deff_K$	3.0	3.0	3.0	3.0	3.0	3.0
Cluster $deff_{KC}$	3.0	4.0	14.9	3.5	3.7	3.0
Park and Lee $deff_{PL}$	1.3	1.4	5.4	1.3	1.4	1.0
Valliant *et al.* $deff_{VDK}$	0.9	1.4	5.2	0.1	1.3	1.1
Proposed						
Exact $deff_{G1}$	0.9	1.5	5.5	0.3	1.5	1.1
Zero-corr. approx. $deff_{G0}$	1.0	1.6	7.6	0.3	1.6	1.1
Equal cluster approx. $deff_{G2}$	3.0	3.9	54.9	25.3	27.7	1.7

* continuous variables; ** categorical variables

The true population design effects, denoted by $deff_{True}$, for the calibration totals range from 0.1 for y_4 to 5.2 for y_3. These are population values, not subject to sampling variability. Notably they are less than one for y_1 and y_4, which means that poststratifying by age and gender will improve estimates of these variables' totals.

The Kish single-stage $deff_K$ is simply $1+\left[CV\left(\mathbf{w}\right)\right]^2$, which is constant regardless of the variable of interest. The cluster sample extension $deff_{KC}$ are all very large, 3.0 or higher. The Park and Lee

$deff_{PL}$'s account for the unequal cluster weights but all exceed one. This also occurred for the approximations, $deff_{G0}$ and $deff_{G2}$. This means for all of these alternatives, the *deff*'s are too large. In particular, one would incorrectly believe that calibration would not be beneficial for y_1 and y_4. The Valliant *et al.* $deff_{VDK}$ equals $deff_{True}$ since they are based on the same formula; the alternative estimator $deff_{G1}$ is very close to $deff_{True}$.

While the Park and Lee $deff_{PL}$ was not designed to capture the gains in calibration, it could by incorporating the k_e-factor used in $deff_{VDK}$. The other components between these two *deff*'s were nearly identical. This is shown in Table 2.

Table 2. Valliant *et al.* and Park and Lee Deff Components, by Variable of Interest

	Variable of Interest					
Design Effect Component	y_1*	y_2*	y_3*	y_4*	Hospital Stay**	Insurance Coverage**
Park and Lee $deff_{PL}$						
W_y^*	0.000	0.000	0.000	0.000	0.000	0.000
τ	0.011	0.017	0.185	0.012	0.015	0.004
$1+(\bar{m}-1)\tau$	0.256	0.413	0.444	0.290	0.350	0.094
Valliant *et al.* $deff_{VDK}$						
δ_e	0.008	0.018	0.200	0.001	0.013	0.004
k_e	0.733	0.980	0.888	0.104	0.964	0.997
$1+(\bar{m}-1)\delta_e$	0.198	0.420	4.802	0.017	0.302	0.091
$k_e\left[1+(\bar{m}-1)\delta_e\right]$	0.145	0.412	4.262	0.002	0.291	0.091

* continuous variables; ** categorical variables

The τ and δ_e values are close, such that differences between these *deff*'s are due to the k_e-factor. While the Park and Lee $deff_{PL}$ was not designed to account for calibration efficiency, Table 2 demonstrates how it can easily be adapted by incorporating k_e.

5. Discussion, Limitations, and Conclusions

We propose new design effects that gauge the impact of calibration weighting adjustments on an estimated total in cluster sampling. Existing design effects include Kish's (1965) "design effect due to weighting" and Park and Lee's (2004) for unequal two-stage cluster weights. Neither of these 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 Park and Lee *deff* does signal whether the HT (or *pwr*) estimator in varying probability sampling is more efficient than *srs*, but does not reflect any gains from using a calibration estimator.

The new design effects introduced in Sec. 3 measure 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 proposed measure $deff_{G1}$ 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. In cases where the auxiliary information is ineffective or is not used, the proposed measure approximates Kish's $deff_K$.

One of our proposed measures, $deff_{G1}$, uses the model underlying the general regression estimator to extend the Spencer measure. The other alternative, $deff_{VDK}$, has a form similar to the familiar $1+\left(\bar{m}-1\right)\rho$ found in many texts. The survey variable, covariates, and weights are required to produce both new design effects. Since the variance (14) is approximately correct in large samples for all calibration estimators, the new design effects should reflect the effects of many forms of commonly used weighting adjustment methods, including poststratification, other forms of the GREG estimator, and raking. 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 *deff*'s do not reflect any gains from those adjustments. The alternative *deff*'s introduced in this paper, thus, serve as a corrective to that deficiency.

References

Brick, M.J., and Montaquila, J. M. (2009). "Nonresponse and weighting" in D. Pfeffermann and C. R. Rao (eds.), *Handbook of Statistics*, *Sample Surveys: Design, Methods and Applications*, **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.

Gabler, S., Haeder, S., and Lahiri, P. (1999). A model-based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology*, **25**, 105-106.

Henry, K., and Valliant, R. (2013). A Design Effect Measure for Calibration Weights in Single-Stage Samples. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, to appear.

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

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

Kish, L. (1987). Weighting in *Deft*^{2}. *The Survey Statistician*, June.

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: Inference and Analysis*, **29B**. Amsterdam: Elsevier BV.

Lynn, P. and Gabler, S. (2005). Approximations to b* in the prediction of Design effects due to clustering, *Survey Methodology*, **31**, 101-104.

Park, I. (2004). Assessing Complex Sample Designs via Design Effect Decompositions. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 4135-4142.

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

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*, Berlin: Springer.

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

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. (2014). `PracTools`: Tools for Designing and Weighting Survey Samples. R package version 0.0-2. http://CRAN.R-project.org/package=PracTools.