Brazilian Journal of Probability and Statistics
2017, Vol. 31, No. 1, 160–178
DOI: 10.1214/16-BJPS307

Calibration estimation of adjusted Kuk's randomized response model for sensitive attribute

Chang-Kyoon Son[a] **and Jong-Min Kim**[b]

[a]*Dongguk University–Gyeongju*
[b]*University of Minnesota–Morris*

Abstract. In this paper, we consider the calibration procedure for Su et al.'s [*Sociol. Methods Res.* **44** (2014) DOI:10.1177/0049124114554459] adjusted Kuk randomized response (RR) technique by using auxiliary information such as gender or age group of respondents associated with the variable of interest. Our proposed calibration method can overcome the problems such as noncoverage and nonresponse. From the efficiency comparison study, we show that the calibrated adjusted Kuk's RR estimators are more efficient than that of Su et al. [*Sociol. Methods Res.* **44** (2014) DOI:10.1177/0049124114554459], when the known population cell and marginal counts of auxiliary information are used for the calibration procedure.

1 Introduction

Warner (1965) first suggested an ingenious survey model called randomized response (RR) technique to procure sensitive information from respondents without disturbing their privacy by using a randomization device which was composed of two questions. One was sensitive and the other was non-sensitive:

Question 1: Do you have a sensitive attribute A? (with probability P),
Question 2: Do you have a non-sensitive attribute $\bar{A}$? (with probability $1-P$).

The probability of a "Yes" answer is given by

$$\phi_W = P\pi + (1-P)(1-\pi). \tag{1.1}$$

Let $n\hat{\phi}_W$ be the number of "Yes" answers in a random sample of n respondents, the estimator $\hat{\pi}_W$ and it's variance $V(\hat{\pi}_W)$ of sensitive proportion π are respectively,

$$\hat{\pi}_W = \frac{\hat{\phi}_W - (1-P)}{2P-1}, \qquad P \neq 1/2, \tag{1.2}$$

$$V(\hat{\pi}_W) = \frac{\pi(1-\pi)}{n} + \frac{P(1-P)}{n(2P-1)^2}. \tag{1.3}$$

Key words and phrases. Randomized response technique, adjusted Kuk's RRT, calibration estimator, post-stratification.

Received January 2015; accepted January 2016.

Kuk (1990) suggested an RR design that made use of two randomization devices. The first randomization device R_1, which is made up a deck of cards each bearing one of two possible questions that has two possible outcomes:

Question 1: Do you have a sensitive attribute A? (with probability θ_1),
Question 2: Do you have a non-sensitive attribute $\bar{A}$? (with probability $1-\theta_1$).

The second randomization device, R_2, which is made up a deck of cards each bearing one of two possible questions that has two possible outcomes:

Question 1: Do you have a non-sensitive attribute $\bar{A}$? (with probability θ_2),
Question 2: Do you have a sensitive attribute A? (with probability $1-\theta_2$).

Assume that a simple random sample with replacement (SRSWR) of size n respondents is selected from the population of interest. Each respondent is to report the first outcome of R_1 if he/she has a sensitive attribute A and the second outcome of R_2 if he/she does not have a sensitive attribute A.

The probability of a "Yes" answer ϕ_K is given by

$$\phi_K = P(Yes) = \pi\theta_1 + (1-\pi)\theta_2. \tag{1.4}$$

Let $n\hat{\phi}_K$ denote the number of "Yes" responses in the sample of size n, the estimator $\hat{\pi}_K$ of π, the proportion of the population in the sensitive group, and it's variance $V(\hat{\pi}_K)$ are given by

$$\hat{\pi}_K = \frac{\hat{\phi}_K - \theta_2}{\theta_1 - \theta_2}, \qquad \theta_1 \neq \theta_2, \tag{1.5}$$

$$V(\hat{\pi}_K) = \frac{\phi_K(1-\phi_K)}{n(\theta_1-\theta_2)^2}. \tag{1.6}$$

Recently, Su et al. (2014) suggested a new RR model compelling answers "Yes" or "No" to each respondent according to his/her selection situation in the randomization device which modified Kuk's randomization device.

It has been a difficult problem for social survey statisticians to deal with nonresponse and noncoverage of survey data. The respondents are unlikely to respond to the survey especially when sensitive questions related to their privacies are asked. In order to adjust the survey nonresponse, we can use auxiliary information to improve the precision of the estimator for the population parameters such as total, mean, and proportion using external data. In terms of calibration procedure, Deville and Särndal (1992), and Deville, Särndal and Sautory (1993) suggested the calibration estimator according to the distance functions.

Tracy et al. (1999) suggested the calibrated estimator of the quantitative RR survey for the quantitative sensitive characteristics, and they suggested the high-order calibration method using the population variance of the auxiliary variable. Recently, Son et al. (2010) suggested the calibrated RR estimators of qualitative

sensitive question survey, and they showed that the calibrated RR estimators are more efficient than that of Waner's and Mangat model.

In this paper, we suggest the calibrated estimator of Su et al. (2014) adjusted Kuk's randomized response technique using auxiliary information such as demographical variables associated with the variable of interest.

In Section 2, we review the adjusted Kuk's RR model suggested by Su et al. (2014). Section 3 proposes the calibration procedure for Su et al.'s RR model, and Section 4 introduces the conditional and unconditional properties of the calibrated RR estimators. Section 5 is devoted to the simulation and a real survey data study in order to compare the efficiencies between the calibrated adjusted Kuk's RR estimators and the original Kuk's RR ones, and Section 6 provides the conclusion.

2 Review of adjusted Kuk's randomized response model

In this section, we review the adjusted Kuk's RR model suggested by Su et al. (2014). Su et al. estimated the proportion of sensitive attribute by suggesting an adjusted Kuk's one. They consider selecting a SRSWR sample of n respondents from the given population of interest. Each respondent in the sample of n respondents is provided with two randomization devices, D_1 and D_2. The randomization device D_1 consists of a deck of cards, each card bearing one of two types of statements: (1) Use randomization device F_1 and (2) use randomization device $\bar{F}_1$ with probabilities θ_1 and $(1-\theta_1)$, respectively. Similarly, the randomization device D_2 consists of a deck of cards, each card bearing one of two statements: (1) Use randomization device F_2 and (2) use randomization device $\bar{F}_2$, with probabilities θ_2 and $(1-\theta_2)$ respectively. Each respondent is instructed to use the first device D_1 if he/she has the sensitive attribute A, and to use the second device D_2 if he/she has the non-sensitive attribute $\bar{A}$.

The device F_1 mentioned by the first outcome of device D_1 consists of two possible mutually exclusive statements: (1) Say "Yes" and (2) say "No" with probabilities P_1 and $(1-P_1)$, respectively. The device $\bar{F}_1$ mentioned by the second outcome of device D_1 also consists of two possible mutually exclusive statements: (1) Say "Yes" and (2) say "No" but with probabilities T_1 and $(1-T_1)$, respectively. Similarly, the device F_2 mentioned by the first outcome of device D_2 consists of two possible mutually exclusive statements: (1) Say "Yes" and (2) say "No" with probabilities P_2 and $(1-P_2)$, respectively. The device $\bar{F}_2$ mentioned by the second outcome of device D_2 also consists of two possible mutually exclusive statements: (1) Say "Yes" and (2) say "No" but with probabilities T_2 and $(1-T_2)$, respectively. A pictorial representation of such a proposed forced randomized response model is given in Figure 1.

In the adjusted Kuk's RR model, the probability of a "Yes" answer is given by:

$$\begin{aligned}\phi &= \pi\big[\theta_1 P_1 + (1-\theta_1)T_1\big] + (1-\pi)\big[\theta_2 P_2 + (1-\theta_2)T_2\big] \\ &= \pi\big[\theta_1(P_1-T_1) - \theta_2(P_2-T_2) + (T_1-T_2)\big] + \theta_2 P_2 + (1-\theta_2)T_2,\end{aligned} \tag{2.1}$$

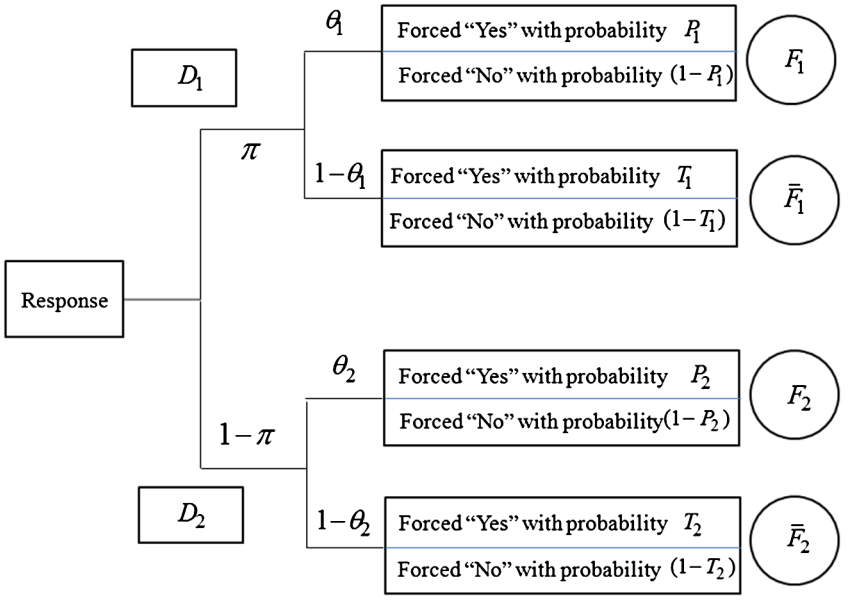

Figure 1 *Adjusted Kuk's forced randomized response model.*

where π is the population proportion of sensitive attribute.

The estimator $\hat{\pi}_s$ of the population proportion of sensitive attribute is

$$\hat{\pi}_s = \frac{\hat{\phi} - \theta_2 P_2 - (1-\theta_2)T_2}{\theta_1(P_1 - T_1) - \theta_2(P_2 - T_2) + (T_1 - T_2)}, \tag{2.2}$$

where $\hat{\phi} = \sum_{k=1}^{n} \frac{y_k}{n}$ is the observed proportion of "Yes" answers in the sample.

The variance of the proposed estimator $\hat{\pi}_s$ is given as follows:

$$V(\hat{\pi}_s) = \frac{\phi(1-\phi)}{n[\theta_1(P_1 - T_1) - \theta_2(P_2 - T_2) + (T_1 - T_2)]^2}. \tag{2.3}$$

If the respondents are selected by simple random sampling without replacement (SRSWOR), then the variance of the proposed estimator $\hat{\pi}_s$ is given as follows:

$$V(\hat{\pi}_s) = \left(\frac{N-n}{N-1}\right) \frac{\phi(1-\phi)}{n[\theta_1(P_1 - T_1) - \theta_2(P_2 - T_2) + (T_1 - T_2)]^2}. \tag{2.4}$$

3 Calibration procedure

The RR survey for sensitive attribute has the limitation to the use of auxiliary information for the privacy protection of respondents. Nevertheless, auxiliary information of respondents of the RR survey may be available some socio-demographical auxiliary information such as gender and age in the population level. In this sec-

tion, we consider the calibration procedure to improve the Su et al.'s RR estimator.

3.1 Known population cell counts

Let y_k be the binomial variable with parameter ϕ. The sample respondents are selected by simple random sampling without replacement (SRSWOR). Then the population proportion reporting "Yes" to RR question is defined by $\bar{y} = N^{-1}\sum_U y_k$ and the counterpart of the sample is $\hat{\bar{y}} = N^{-1}\sum_s d_k y_k$, where $d_k = 1/\upsilon_k$ is the sampling design weight. The auxiliary information $\tau_x = \sum_{k\in U} x_k$ is given in the form of known cell counts in contingency table with two dimensions as follows:

$$\sum_{k\in U} x_k = (N_{11}, N_{12}, \ldots, N_{ij}, \ldots, N_{rc}). \tag{3.1}$$

For Su et al.'s RR model, the sample proportion of answering "Yes", $\bar{y}$, can be rewritten as follows:

$$\bar{y} = \frac{1}{N}\sum_{k=1}^{n} d_k y_k, \tag{3.2}$$

where the original sampling weight is $d_k = N/n$ for SRSWOR.

The original sampling weight d_k is replaced by the new weight $w_k = d_k N_{ij}/\hat{N}_{ij}$, and then the calibrated sample proportion $\bar{y}$ is given by

$$\bar{y}_{post} = \frac{1}{N}\sum_{k=1}^{n} w_k y_k = \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{c} N_{ij}\tilde{y}_{ij}, \tag{3.3}$$

where $\tilde{y}_{ij} = \sum_{k=1}^{n_{ij}} d_k y_k/\hat{N}_{ij}$ is the weighted proportion in the sample cell with $\hat{N}_{ij} = \sum_{s_{ij}} d_k$.

Theorem 3.1. *If the respondents are selected by SRSWOR, $\hat{N}_{ij} = d_k n_{ij} = (N/n)n_{ij}$, then the post-stratified Su et al.'s RR estimator is given by*

$$\hat{\pi}_{post} = \sum_{i=1}^{r}\sum_{j=1}^{c}\left(\frac{N_{ij}}{N}\right)\frac{\hat{\phi}_{ij} - \theta_2 P_2 - (1-\theta_2)T_2}{\theta_1(P_1 - T_1) - \theta_2(P_2 - T_2) + (T_1 - T_2)}, \tag{3.4}$$

where $\hat{\phi}_{ij} = \sum_k \frac{y_k}{n_{ij}}$ is the observed proportion of "Yes" answers in the sample cell (i, j).

Proof. From (2.2) the Su et al.'s RR estimator $\hat{\pi}_c$, we can rewrite a sample proportion as (3.3) under SRSWOR and then the post-stratified RR estimator is given

by

$$\begin{aligned}\hat{\pi}_{post} &= \frac{1}{N}\sum_{k=1}^{n}\frac{d_k y_k - \theta_2 P_2 - (1-\theta_2)T_2}{\theta_1(P_1-T_1)-\theta_2(P_2-T_2)+(T_1-T_2)}\\ &= \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{c} N_{ij}\frac{\sum_{k=1}^{n_{ij}} d_k \tilde{y}_k - \theta_2 P_2 - (1-\theta_2)T_2}{\theta_1(P_1-T_1)-\theta_2(P_2-T_2)+(T_1-T_2)}\\ &= \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{N_{ij}}{N}\frac{\sum_{k=1}^{n_{ij}} y_k/n_{ij} - \theta_2 P_2 - (1-\theta_2)T_2}{\theta_1(P_1-T_1)-\theta_2(P_2-T_2)+(T_1-T_2)}\\ &= \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{N_{ij}}{N}\frac{\hat{\phi}_{ij} - \theta_2 P_2 - (1-\theta_2)T_2}{\theta_1(P_1-T_1)-\theta_2(P_2-T_2)+(T_1-T_2)}.\end{aligned}$$

□

3.2 The only known population marginal counts

We consider using the knowledge of population cell counts of the auxiliary variable in the previous calibration procedure. But if we only know the population marginal counts from auxiliary information, we can use the knowledge of marginal counts in calibration procedure as the following,

$$\sum_{k\in U} x_k = (N_{1+}, N_{2+}, \ldots, N_{r+}, N_{+1}, N_{+2}, \ldots, N_{+c})', \tag{3.5}$$

where $N_{i+} = \sum_{j=1}^{c} N_{ij}$, $N_{+j} = \sum_{i=1}^{r} N_{ij}$.

From (3.5), we define the auxiliary variable vector $x_k = (\delta_{1\cdot k}, \ldots, \delta_{r\cdot k}, \delta_{\cdot 1k}, \ldots, \delta_{\cdot ck})'$, where $\delta_{i\cdot k} = 1$, if the respondent k is in row i and 0 otherwise, $\delta_{\cdot jk} = 1$ if the respondent k is in column j and 0 otherwise.

We denote the Lagrange multiplier as $\varphi = (u_1, \ldots, u_r, v_1, \ldots, v_c)'$ so that we can express $x_k'\varphi = u_i + v_j$, which can be written as $F(x_k'\varphi) = F(u_i + v_j)$, where $F = (\partial G/\partial w)^{-1}$ is defined as Deville and Särndal (1992). The calibration equations are

$$\sum_{j=1}^{c} \hat{N}_{ij} F(u_i + v_j) = N_{i+} \qquad \text{for } i = 1, 2, \ldots, r, \tag{3.6}$$

$$\sum_{i=1}^{r} \hat{N}_{ij} F(u_i + v_j) = N_{+j} \qquad \text{for } j = 1, 2, \ldots, c, \tag{3.7}$$

where u_i and v_j are determined by iterative computation.

We can obtain the calibrated cell counts estimate $\hat{N}_{ij}^{w} = \hat{N}_{ij} F(u_i + v_j)$, and then the calibrated weight is $w_k = d_k \hat{N}_{ij}^{w}/\hat{N}_{ij}$. As a result the calibration estimator for population proportion ϕ is given by

$$\bar{y}_{cal} = \frac{1}{N}\sum_{k=1}^{n} w_k y_k = \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{c} \hat{N}_{ij}^{w}\tilde{y}_{ij}. \tag{3.8}$$

Theorem 3.2. *By* (3.8), *if the respondents are selected by SRSWOR, then the calibrated Su et al.'s RR estimator is given by*

$$\hat{\pi}_{cal} = \sum_{i=1}^{r}\sum_{j=1}^{c}\left(\frac{\hat{N}_{ij}^{w}}{N}\right)\frac{\hat{\phi}_{ij} - \theta_2 P_2 - (1-\theta_2)T_2}{\theta_1(P_1 - T_1) - \theta_2(P_2 - T_2) + (T_1 - T_2)}. \tag{3.9}$$

Proof. Refer to the proof of Theorem 3.1. □

4 Variances and its estimator of calibrated Su et al.'s RR estimators

In this section, we investigate the conditional and unconditional properties of the calibrated Su et al.'s RR estimator. The conditional variance given the cell or marginal count of population, $V(\cdot|\hat{N})$, can be derived from the cell or marginal information of population level, and the unconditional variance is derived from the double expectation of estimates. In addition, we derive the variance estimator of the proposed calibration RR estimator.

4.1 Conditional variances

We consider a row effect and a column effect in two-way contingency table for RR survey data. Let the two cross effect factors explain the population proportion reporting "Yes" for RR questions, then we parameterize the finite population using the ANOVA representing that for respondent k in population cell U_{ij}, $y_k = \alpha_i + \beta_j + E_k$, where y_k is the binomial variable to RR question. If α_i is a row effect, β_j a column effect, and E_k is an error term, then α_i and β_j are fixed unknown values defined by calibration equations

$$\sum_{j=1}^{c} N_{ij}(\alpha_i + \beta_j) = N_{i+}\phi_{i+} \qquad \text{for } i = 1, 2, \ldots, r, \tag{4.1}$$

$$\sum_{i=1}^{r} N_{ij}(\alpha_i + \beta_j) = N_{+j}\phi_{+j} \qquad \text{for } j = 1, 2, \ldots, c. \tag{4.2}$$

Let us decompose the kth error term $E_k = L_{ij} + R_k$, where $L_{ij} = \phi_{ij} - (\alpha_i + \beta_j)$ is an interaction term, and $R_k = y_k - \phi_{ij}$ is the deviation from $\phi_{ij} = \sum_{U_{ij}} y_{ij}/N_{ij}$, where ϕ_{ij} represents the population proportion of "Yes" to the RR question in cell ij. The restrictions for the interaction term are

$$\sum_{i=1}^{r} N_{ij}L_{ij} = \sum_{j=1}^{c} N_{ij}L_{ij} = 0. \tag{4.3}$$

The variable of interest y_k can be written as $y_k = \alpha_i + \beta_j + L_{ij} + R_k$, so that the calibrated Su et al.'s RR estimator can be expressed by

$$\bar{y}_{cal} = \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{c}\hat{N}_{ij}^{w}(\alpha_i + \beta_j + L_{ij} + \tilde{R}_{ij}), \tag{4.4}$$

where $\tilde{R}_{ij} = \sum_{k=1}^{n_{ij}} d_k R_k / \hat{N}_{ij}$ are the deviation proportion of sample cells and $\hat{N}_{ij}^w$ are the calibrated cell counts.

Also, we can express the calibration equation of y_k as follows:

$$\frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{c} N_{ij}(\alpha_i + \beta_j) = \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{c} \hat{N}_{ij}^w(\alpha_i + \beta_j). \tag{4.5}$$

Let the population proportion answering "Yes" for the RR question, $\phi = N^{-1}\sum_U y_k$ be denoted by the left-hand side of equation (4.5), then we can express the error of $\bar{y}_{cal}$ as

$$\bar{y}_{cal} - \phi = \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{c}(\hat{N}_{ij}^w - N_{ij})L_{ij} + \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{c}\hat{N}_{ij}^w\tilde{R}_{ij}. \tag{4.6}$$

Similar to (4.6), the error of the post-stratified estimator $\bar{y}_{post}$ is

$$\bar{y}_{post} - \phi = \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{c} N_{ij}\tilde{R}_{ij}. \tag{4.7}$$

The conditional biases $B_c = B(\cdot|\hat{N})$ of the estimators of population means, $\hat{\pi}_{post}$ and $\hat{\pi}_{cal}$, can be expressed by

$$\begin{aligned} B_c(\hat{\pi}_{post}) = &\left(\frac{1}{\theta_1(P_1 - T_1) - \theta_2(P_2 - T_2) + (T_1 - T_2)}\right) \\ &\times \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{c} N_{ij}E_c(\tilde{R}_{ij}), \end{aligned} \tag{4.8}$$

$$\begin{aligned} B_c(\hat{\pi}_{cal}) = &\left(\frac{1}{\theta_1(P_1 - T_1) - \theta_2(P_2 - T_2) + (T_1 - T_2)}\right) \\ &\times \left[\frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{c}(\hat{N}_{ij}^w - N_{ij})L_{ij} + \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{c}\hat{N}_{ij}^w E_c(\tilde{R}_{ij})\right], \end{aligned} \tag{4.9}$$

respectively.

From (4.8) and (4.9), the conditional expectation is $E_c(\tilde{R}_{ij}) = 0$ or nearly 0 for all i, j, because the sampling design is SRSWOR. Then the inclusion probability υ_k is constant. The conditional bias of post-stratified estimator $B_c(\hat{\pi}_{post}) \approx 0$, whereas $B_c(\hat{\pi}_{cal}) = N^{-1}\sum_{i=1}^{r}\sum_{j=1}^{c}(\hat{N}_{ij}^w - N_{ij})L_{ij}$. For a large sample, $\hat{N}_{ij}^w$ is closed to N_{ij}, and then the conditional bias of $\bar{y}_{cal}$ is asymptotically equal to that of $\bar{y}_{post}$.

The conditional variance of the post-stratified Su et al.'s RR estimator $\hat{\pi}_{post}$ can be rewritten by

$$V_c(\hat{\pi}_{post}) = \sum_{i=1}^{r}\sum_{j=1}^{c}\left(\frac{N_{ij}}{N}\right)^2\left[\left(\frac{N_{ij}-n_{ij}}{N_{ij}-1}\right) \times \frac{\phi_{ij}(1-\phi_{ij})}{n_{ij}\{\theta_1(P_1-T_1)-\theta_2(P_2-T_2)+(T_1-T_2)\}^2}\right]. \tag{4.10}$$

Also, the conditional variance of calibration Su et al.'s RR estimator $\hat{\pi}_{cal}$ is

$$V_c(\hat{\pi}_{cal}) = \sum_{i=1}^{r}\sum_{j=1}^{c}\left(\frac{\hat{N}_{ij}^{w}}{N}\right)^2\left[\left(\frac{N_{ij}-n_{ij}}{N_{ij}-1}\right) \times \frac{\phi_{ij}(1-\phi_{ij})}{n_{ij}\{\theta_1(P_1-T_1)-\theta_2(P_2-T_2)+(T_1-T_2)\}^2}\right]. \tag{4.11}$$

If the interaction terms L_{ij} are negligible in (4.9), then the conditional variances of $\hat{\pi}_{cal}$ are equal to the conditional variances of $\hat{\pi}_{post}$ replacing $\hat{N}_{ij}^{w}$ by N_{ij}. Ordinarily, it is reasonable that the conditional variances of the calibration estimators are larger than the conditional variances of the post-stratified estimators. Also, we note the conditional bias of the post-stratified estimators are unaffected by interaction, whereas that of the calibration estimators depend on interaction.

4.2 Unconditional variances

The unconditional variance is $V(\cdot) = E(V_c) + V(B_c)$, we can derive the unconditional variances of calibrated Su et al.'s RR estimators $\hat{\pi}_{post}$ and $\hat{\pi}_{cal}$.

Theorem 4.1. *The unconditional variance of the post-stratified Su et al.'s RR estimator can be expressed by*

$$\begin{aligned} V(\hat{\pi}_{post}) = &\sum_{i=1}^{r}\sum_{j=1}^{c}\left(\frac{N_{ij}}{N}\right) \\ &\times\left[(1-f)\frac{\phi_{ij}(1-\phi_{ij})}{n\{\theta_1(P_1-T_1)-\theta_2(P_2-T_2)+(T_1-T_2)\}^2}\right] \\ &+\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}\left(1-\frac{N_{ij}}{N}\right) \\ &\times\left[(1-f)\frac{\phi_{ij}(1-\phi_{ij})}{n\{\theta_1(P_1-T_1)-\theta_2(P_2-T_2)+(T_1-T_2)\}^2}\right]. \end{aligned} \tag{4.12}$$

Proof. By Cochran (1977), the size of sample cell n_{ij} is the random variable with $E(n_{ij}) = n(N_{ij}/N)$, $V(n_{ij}) = nN_{ij}/N(1 - N_{ij}/N)$ for the post-stratification. n_{ij} can be expressed by

$$n_{ij} = n\frac{N_{ij}}{N}\left(1 - \frac{n(N_{ij}/N) - n_{ij}}{n(N_{ij}/N)}\right). \tag{4.13}$$

Thus the $1/n_{ij}$ can be written

$$\frac{1}{n_{ij}} = \frac{1}{n(N_{ij}/N)}\left(1 - \frac{n(N_{ij}/N) - n_{ij}}{n(N_{ij}/N)} + \frac{(nN_{ij}/N - n_{ij})^2}{(nN_{ij}/N)^2} - \cdots\right).$$

Then the expectation of $1/n_{ij}$ is

$$\begin{aligned} E\left(\frac{1}{n_{ij}}\right) &\cong \frac{1}{n(N_{ij}/N)} + \frac{n(N_{ij}/N)(1 - N_{ij}/N)}{(nN_{ij}/N)^2} \\ &= \frac{1}{n(N_{ij}/N)} + \frac{(1 - N_{ij}/N)}{(nN_{ij}/N)^2}. \end{aligned}$$

Substitute $E(n_{ij}) = n(N_{ij}/N)$ and (4.13) into the expectation of (4.10), and after some algebra, we can obtain (4.12). □

Theorem 4.2. *The unconditional variance of calibrated Su et al.'s RR estimator is given by*

$$\begin{aligned} V(\hat{\pi}_{cal}) = &\sum_{i=1}^{r}\sum_{j=1}^{c}\left(\frac{\hat{N}_{ij}^{w}}{N}\right) \\ &\times \left[\frac{\phi_{ij}(1 - \phi_{ij})}{n\{\theta_1(P_1 - T_1) - \theta_2(P_2 - T_2) + (T_1 - T_2)\}^2}(1 - f)\right] \\ &+ \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}\left(1 - \frac{\hat{N}_{ij}^{w}}{N}\right) \\ &\times \left[\frac{\phi_{ij}(1 - \phi_{ij})}{n\{\theta_1(P_1 - T_1) - \theta_2(P_2 - T_2) + (T_1 - T_2)\}^2}(1 - f)\right] \\ &+ \frac{1 - f}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}\left(\frac{\hat{N}_{ij}^{w}}{N}\right)^2 \\ &\times \frac{L_{ij}^2}{n\{\theta_1(P_1 - T_1) - \theta_2(P_2 - T_2) + (T_1 - T_2)\}^2}. \end{aligned} \tag{4.14}$$

Proof. We can obtain the first and second terms of right-hand side in (4.14) from Theorem 4.1 replacing $\hat{N}_{ij}^{w}$ by N_{ij}. For the third term of (4.14), the variance of

conditional bias becomes $V(B_c(\hat{\pi}_{cal})) = V(\sum_i \sum_j \hat{N}_{ij}^w L_{ij})$ from (4.9) as follows:

$$\begin{aligned} V\big(B_c(\hat{\pi}_{cal})\big) &= \left(\frac{1}{\theta_1(P_1 - T_1) - \theta_2(P_2 - T_2) + (T_1 - T_2)}\right)^2 \\ &\quad \times V\left[\frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{c}(\hat{N}_{ij}^w - N_{ij})L_{ij}\right] \\ &= \left(\frac{1}{\theta_1(P_1 - T_1) - \theta_2(P_2 - T_2) + (T_1 - T_2)}\right)^2 \\ &\quad \times V\left[\frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{c}\hat{N}_{ij}^w L_{ij}\right] \\ &= \left(\frac{1}{\theta_1(P_1 - T_1) - \theta_2(P_2 - T_2) + (T_1 - T_2)}\right)^2 \\ &\quad \times \sum_{i=1}^{r}\sum_{j=1}^{c}\left(\frac{\hat{N}_{ij}^w}{N}\right)^2\left(\frac{1-f}{n}\right)L_{ij}^2. \end{aligned}$$

□

From the unconditional variances of the calibrated Su et al.'s RR estimator (4.14), the first term of the unconditional variances equals the post-stratified variance replacing $\hat{N}_{ij}^w$ by N_{ij}. Therefore, if $E(\hat{N}_{ij}^w) \cong N_{ij}$ for large sample, then the last terms of the (4.14) is negligible. Hence, the unconditional variance of the calibrated Poisson RR estimator equals to that of the post-stratified Su et al.'s RR estimator.

4.3 Variance estimation

The variance estimator of calibrated Sue et al. RR estimator can be derived from the calibration procedure. In Section 4.1, we assumed the two-way ANOVA model as $y_k = \alpha_i + \beta_j + E_k$ in population level. Then we can consider sample level model as $y_k = \hat{\alpha}_i + \hat{\beta}_j + e_k$. The variance estimator is calculated from the sample-based calibration equations as follows:

$$\sum_{j=1}^{c}\hat{N}_{ij}^w(\hat{\alpha}_i + \hat{\beta}_j) = \sum_{j=1}^{c}\hat{N}_{ij}^w\hat{\phi}_{ij} \qquad \text{for } i = 1, 2, \ldots, r, \tag{4.15}$$

$$\sum_{i=1}^{r}\hat{N}_{ij}^w(\hat{\alpha}_i + \hat{\beta}_j) = \sum_{i=1}^{r}\hat{N}_{ij}^w\hat{\phi}_{ij} \qquad \text{for } j = 1, 2, \ldots, c. \tag{4.16}$$

For SRSWOR, the variance estimator of the calibration Su et al. RR estimator is given by

$$\begin{aligned}\hat{V}(\hat{\pi}_{cal}) = {} & \frac{n(1-f)}{n-1}\sum_{i=1}^{r}\sum_{j=1}^{c}\left(\frac{\hat{N}_{ij}^{w}}{N}\right) \\ & \times\left[\frac{\hat{\phi}_{ij}(1-\hat{\phi}_{ij})}{n_{ij}\{\theta_1(P_1-T_1)-\theta_2(P_2-T_2)+(T_1-T_2)\}^2}\right] \\ & + \frac{n(1-f)}{n-1}\sum_{i=1}^{r}\sum_{j=1}^{c}\left(1-\frac{\hat{N}_{ij}^{w}}{N}\right) \\ & \times\left[\frac{\hat{\phi}_{ij}(1-\hat{\phi}_{ij})}{n_{ij}\{\theta_1(P_1-T_1)-\theta_2(P_2-T_2)+(T_1-T_2)\}^2}\right] \\ & + \frac{n(1-f)}{n-1}\sum_{i=1}^{r}\sum_{j=1}^{c}\left(\frac{\hat{N}_{ij}^{w}}{N}\right)^2 \\ & \times\frac{\hat{L}_{ij}^{2}}{n_{ij}\{\theta_1(P_1-T_1)-\theta_2(P_2-T_2)+(T_1-T_2)\}^2}.\end{aligned} \tag{4.17}$$

5 Efficiency comparison study

5.1 Numerical comparison

We assume the population with rows and columns in contingency table according to auxiliary variables with 2×2 dimensions. As discussed in Deville et al. (1993), this dimension of the population and sample contingency table can be extended to more than 2×2 dimensions. We generate a population with size N (=10,000), and then it classifies with 2×2 table according to size of random generated number.

Table 1 shows the population distribution of the respondents, each cell count denoted by N_{ij}, which can be known from the socio-demographic information for respondents. Let N_{i+} and N_{+j} denote the row and column marginal counts, respectively. If the population cell counts N_{ij} are known, then we can use the post-stratified estimator, and if these counts are unknown but the marginal counts N_{i+} and N_{+j} are known, then we can use the calibration estimator.

Table 2 describes the sample distribution of the respondent selected by SRSWOR with size of n (=1000) and each cell count n_{ij} observed from the survey.

We obtain the response set of size 200 from the sample Table 2 according to a given sample proportion $\bar{y}$ of reporting "Yes" to a sensitive attribute as followed by Table 3.

As a result, we calibrate the proportion of respondents reporting "Yes" in sample cells according to the available information N_{ij} or N_{i+} and N_{+j}. We compute the

Table 1 *Population distribution*

	Gender		
Dwelling area	Male	Female	Total
Urban	N_{11} (=3711)	N_{12} (=1257)	N_{1+} (=4968)
Rural	N_{21} (=1296)	N_{22} (=3736)	N_{2+} (=5032)
Total	N_{+1} (=5007)	N_{+2} (=4993)	N (=10,000)

Table 2 *Sample distribution*

	Gender		
Dwelling area	Male	Female	Total
Urban	n_{11} (=376)	n_{12} (=139)	n_{1+} (=515)
Rural	n_{21} (=127)	n_{22} (=358)	n_{2+} (=485)
Total	n_{+1} (=503)	n_{+2} (=497)	n (=1000)

Table 3 *Respondents distribution*

	Gender		
Dwelling area	Male	Female	Total
Urban	78	25	103
Rural	33	64	97
Total	111	89	200

unconditional variance of calibration and ordinary Su et al.'s RR model changing the population proportion π for sensitive attribute and the selection probabilities P_1, P_2, T_1 and T_2. We compare the relative efficiencies (RE) between the unconditional variance of the calibration Su et al.'s RR estimator as follows:

$$RE = \frac{V(\hat{\pi}_s)}{V(\hat{\pi}_{cal})},$$

where $V(\hat{\pi}_{cal})$ represents the variance of post-stratified and calibrated estimator.

From Table 4 and Table 5, we found that the post-stratified Su et al.'s RR estimator is more efficient than of original Su et al.'s RR estimator. When a population proportion of a sensitive attribute is small, that is less equal than 0.4, then the post-stratified estimator is more efficient. But if a population proportion of an sensitive attribute is greater than or equal to 0.6 and selection probabilities of RR question P_1, T_1 and T_2 are over 0.8, then the RE of our post-stratified estimator is less than 1.

Table 4 *Relative efficiencies of post-stratified Su et al.'s estimator*

	π ($P_2 = 0.5$, $T_1 = 0.6$, $T_2 = 0.7$)			
P_1	0.1	0.2	0.3	0.4
0.1	2.1985	2.4344	2.6654	2.8916
0.2	2.1569	2.3529	2.5455	2.7348
0.3	2.1152	2.2708	2.4243	2.5756
0.4	2.0734	2.1881	2.3017	2.4141
	($P_2 = 0.6$, $T_1 = 0.7$, $T_2 = 0.8$)			
0.1	1.7111	1.9997	2.2811	2.5555
0.2	1.6677	1.9150	2.1569	2.3937
0.3	1.6242	1.8296	2.0313	2.2295
0.4	1.5805	1.7436	1.9043	2.0629
	($P_2 = 0.7$, $T_1 = 0.8$, $T_2 = 0.9$)			
0.1	1.2020	1.5476	1.8830	2.2088
0.2	1.1566	1.4594	1.7543	2.0419
0.3	1.1111	1.3705	1.6242	1.8724
0.4	1.0654	1.2809	1.4925	1.7003

Table 5 *Relative efficiencies of post-stratified Su et al.'s estimator*

	π ($P_2 = 0.1$, $T_1 = 0.6$, $T_2 = 0.7$)				
P_1	0.5	0.6	0.7	0.8	0.9
0.5	2.5254	2.5555	2.5856	2.6156	2.6455
0.6	2.3222	2.3119	2.3017	2.2914	2.2811
0.7	2.1152	2.0629	2.0103	1.9574	1.9043
0.8	1.9043	1.8081	1.7111	1.6133	1.5146
0.9	1.6894	1.5476	1.4039	1.2584	1.1111
	($P_2 = 0.2$, $T_1 = 0.7$, $T_2 = 0.8$)				
0.5	2.2192	2.2914	2.3631	2.4344	2.5052
0.6	2.0103	2.0419	2.0734	2.1048	2.1361
0.7	1.7974	1.7867	1.7759	1.7651	1.7543
0.8	1.5805	1.5256	1.4704	1.4150	1.3593
0.9	1.3593	1.2584	1.1566	1.0539	0.9503
	($P_2 = 0.3$, $T_1 = 0.8$, $T_2 = 0.9$)				
0.5	1.9043	2.0208	2.1361	2.2502	2.3631
0.6	1.6894	1.7651	1.8403	1.9150	1.9892
0.7	1.4704	1.5035	1.5366	1.5695	1.6023
0.8	1.2471	1.2359	1.2246	1.2133	1.2020
0.9	1.0195	0.9619	0.9040	0.8458	0.7873

Table 6 *Relative efficiencies of calibrated Su et al.'s estimator*

	π ($P_2 = 0.5$, $T_1 = 0.6$, $T_2 = 0.7$)			
P_1	0.1	0.2	0.3	0.4
0.1	2.1055	2.3390	2.5689	2.7954
0.2	2.0646	2.2582	2.4494	2.6382
0.3	2.0235	2.1770	2.3289	2.4794
0.4	1.9823	2.0953	2.2075	2.3188
	($P_2 = 0.6$, $T_1 = 0.7$, $T_2 = 0.8$)			
0.1	1.6279	1.9100	2.1871	2.4594
0.2	1.5857	1.8269	2.0646	2.2986
0.3	1.5433	1.7434	1.9410	2.1362
0.4	1.5009	1.6595	1.8165	1.9720
	($P_2 = 0.7$, $T_1 = 0.8$, $T_2 = 0.9$)			
0.1	1.1356	1.4690	1.7957	2.1158
0.2	1.0920	1.3836	1.6700	1.9514
0.3	1.0484	1.2978	1.5433	1.7852
0.4	1.0046	1.2114	1.4157	1.6174

Similar as the post-stratified estimator, we can show that the calibrated Su et al.'s estimator is more efficient than the original Su et al.'s estimator in Table 6 and Table 7. The RE of calibration estimator is less than that of the post-stratified estimator because the former uses the marginal information in the weighting adjustment procedure, and on the contrary the latter uses the cell information of population level. From Table 7, when the population proportion of sensitive attribute is over 0.5 and the selection probabilities P_1, T_1 and T_2 are over 0.8, we can find that the RE of calibration estimator is less than 1. When the selection probabilities of RR question P_1, P_2, T_1 and T_2 are increasing to 0.9 then the efficiency of proposed calibration estimator is decreasing. As a result, our proposed calibration estimator is more efficient than the Su et al.'s RR estimator except in the case of the large value of proportion of sensitive attribute. It means that the calibration RR estimator which uses auxiliary information of respondent such as socio-demographic variables, gender, age group or dwelling area can improve the original RR estimator although the available information is limited to protect the respondent privacy.

5.2 Comparison for real survey data

In this section, we consider the proposed estimation method using the post-stratification and calibration with real survey data. We obtained data from undergraduate students (50 for freshman and sophomore years) in the Department of Applied Statistics at Dongguk University in Gyeongju. Table 8 shows the population and sample distribution according to gender and grade/year.

Table 7 *Relative efficiencies of calibrated Su et al.'s estimator*

	π ($P_2 = 0.1$, $T_1 = 0.6$, $T_2 = 0.7$)				
P_1	0.5	0.6	0.7	0.8	0.9
0.5	2.1260	2.1973	2.2683	2.3390	2.4093
0.6	1.9204	1.9514	1.9823	2.0132	2.0440
0.7	1.7120	1.7015	1.6910	1.6805	1.6700
0.8	1.5009	1.4477	1.3943	1.3407	1.2870
0.9	1.2870	1.1898	1.0920	0.9937	0.8947
	($P_2 = 0.2$, $T_1 = 0.7$, $T_2 = 0.8$)				
0.5	2.1260	2.1973	2.2683	2.3390	2.4093
0.6	1.9204	1.9514	1.9823	2.0132	2.0440
0.7	1.7120	1.7015	1.6910	1.6805	1.6700
0.8	1.5009	1.4477	1.3943	1.3407	1.2870
0.9	1.2870	1.1898	1.0920	0.9937	0.8947
	($P_2 = 0.3$, $T_1 = 0.8$, $T_2 = 0.9$)				
0.5	1.8165	1.9307	2.0440	2.1566	2.2683
0.6	1.6068	1.6805	1.7539	1.8269	1.8996
0.7	1.3943	1.4264	1.4583	1.4903	1.5221
0.8	1.1790	1.1681	1.1573	1.1464	1.1356
0.9	0.9608	0.9057	0.8505	0.7952	0.7396

Table 8 *Population and sample distributions*

	Population			Sample		
	Gender			Gender		
Grade	Male	Female	Total	Male	Female	Total
Freshman	N_{11} (=22)	N_{12} (=14)	N_{1+} (=36)	n_{11} (=21)	n_{12} (=11)	n_{1+} (=32)
Sophomore	N_{21} (=15)	N_{22} (=9)	N_{2+} (=24)	n_{21} (=12)	n_{22} (=6)	n_{2+} (=18)
Total	N_{+1} (=37)	N_{+2} (=23)	N (=60)	n_{+1} (=33)	n_{+2} (=17)	n (=50)

From Su et al. model with probabilities $\theta_1 = 0.2$, $P_1 = 1/6$ and $T_1 = 2/6$ for randomization device D_1 and $\theta_2 = 0.3$, $P_2 = 4/6$ and $T_2 = 2/6$ for randomization device D_2, respectively. In order to answer the question, the respondents used the mobile phone two apps with spindle having 0∼9 score to determine probabilities θ_1 and θ_2 and dice to determine P_1, T_1, P_2 and T_2. We obtain the response set of size 50 from Table 8 according to a given sample proportion $\bar{y}$ of reporting "Yes" to a sensitive attribute as followed by Table 9 using the randomized response questionnaire in the Appendix.

Table 9 *Respondents distribution*

	Gender		
Grade	Male	Female	Total
Freshman	7	7	14
Sophomore	2	3	5
Total	16	10	19

Table 10 *Estimation results*

Methods	Estimated proportions	Stderr	RE
Directed question ($\hat{\pi}_D$)	0.14	0.048494	–
Su et al. model ($\hat{\pi}_{Su}$)	0.40	0.029975	–
Post_Su et al. ($\hat{\pi}_{post}$)	0.41	0.029106	1.029845
Cal_Su et al. ($\hat{\pi}_{cal}$)	0.41	0.032186	0.931278

To compare efficiency between the directed question and randomized response model, we use the directed question and Su et al. model for the same group. The students answer the following for the directed question:

Question: *Have you ever felt the sexual impulses to men or women in your class?*

In Table 10, we obtain the estimates from survey $\hat{\pi}_D = 0.14$ for the directed question, $\hat{\pi}_{Su} = 0.4$ for the Su et al. model, $\hat{\pi}_{post} = 0.41$ for post stratified Su et al. and $\hat{\pi}_{cal} = 0.41$ for the calibrated Su et al. model, respectively.

The relative efficiency is 1.029 and 0.93 for the post stratified and calibrated estimator so that the post stratified estimator is more efficient than Su et al. model. As from Table 4 to 7, the proposed estimates are less than 1 for some probabilities P_1, T_1, P_2 and T_2. We find that the post-stratified estimator is more efficient than the Su et al. model but the calibrated estimator is not.

6 Concluding remarks

This paper considered the calibration procedure to reduce the variances of estimators for Su et al. (2014) which adjusted Kuk's RR estimator. Although the RR survey has a limitation of using auxiliary information for a privacy protection of respondents, we can use any auxiliary variable for respondents such as socio-demographic variable. In this respect, we suggest the calibrated Su et al.'s RR estimator to improve nonresponse and noncoverage.

From the simulation study to compare the proposed and ordinary estimator, we find that the suggested estimators are more efficient than the existing ordinary Su

et al.'s RR estimator. And from real survey data we find that the Su et al.'s RR estimator is higher than the directed question estimator for the sensitive attribute. Also, our proposed estimator is little higher than Su et al.'s model and the efficiency of the post-stratified estimator is greater than it.

Appendix

Randomization device [D1]

☞ Using the spindle app in the mobile phone of the respondents when the outcomes are "0" or "1", goes to device <F1>, otherwise goes to <F1*>

<F1>
Q: ***Have you ever felt the sexual impulses to men or women in your class?***
☞ Response—use the dice app in the mobile phone of the respondents
When the outcome of dice "1", forced answer "yes"

<F1*>
Q: ***Have you never felt the sexual impulses to men or women in your class?***
☞ Response—use the dice app in the mobile phone of the respondents
When the outcome of dice "1"or "2", forced answer "yes"

Randomization device [D2]

☞ Using the spindle app in the mobile phone of the respondents "0", "1" or "2" goes to device <F2>, otherwise goes to <F2*>

<F2>
Q: ***Have you never felt the sexual impulses to men or women in your class?***
☞ Response—use the dice app in the mobile phone of the respondents
When the outcome of dice "3", "4", "5" or "6", forced answer "yes"

<F2*>
Q: ***Have you ever felt the sexual impulses to men or women in your class?***
☞ Response—use the dice app in the mobile phone of the respondents "1" or "5"
When the outcome of dice "1" or "5", forced answer "yes"

Acknowledgment

This work was supported by the Dongguk University Research Fund of 2012.

References

Cochran, W. G. (1977). *Sampling Techniques*, 3rd ed. New York: John Wiley and Sons. MR0474575

Deville, J. C. and Särndal, C. E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.* **87**, 376–382. MR1173804

Deville, J. C., Särndal, C. E. and Sautory, O. (1993). Generalized ranking procedures in survey sampling. *J. Amer. Statist. Assoc.* **88**, 1013–1020.

Kuk, A. Y. C. (1990). Asking sensitive questions indirectly. *Biomerika* **77**, 436–438. MR1064822

Son, C. K., Hong, K. H., Lee, G. S. and Kim, J. M. (2010). The calibration for randomized response estimator. *Comm. Statist. Theory Methods* **39**, 3163–3177. MR2755431

Su, S. C., Sedory, S. A. and Singh, S. (2014). Kuk's model adjusted for protection and efficiency. *Sociol. Methods Res.* DOI:10.1177/0049124114554459.

Tracy, D. S., Singh, H. E. and Singh, R. (1999). Constructing an unbiased estimator of population mean in finite populations using auxiliary information. *Statist. Papers* **40**, 363–368. MR1716499

Warner, S. L. (1965). Randomized response; a survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.* **60**, 63–69.

Department of Applied Statistics
Dongguk University–Gyeongju
Dong-daero 123
Gyeongju, Gyeonbuk 780-714
Korea
E-mail: ckson85@dongguk.ac.kr

Statistics Discipline
Division of Science and Mathematics
University of Minnesota–Morris
Morris, Minnesota 56267
USA
E-mail: jongmink@morris.umn.edu