Research Article GU J Sci 32(1): 372-383 (2019)

Gazi University

Journal of Science

http://dergipark.gov.tr/gujs

On Generalized Additive Scrambled Response Modeling in Sensitive Surveys

Waqas ARSHAD[1,*], Zawar HUSSAIN[1]

[1]*Department of Statistics, Quaid-i-Azam University 45320, Islamabd, Pakistan*

Article Info

Received: 29/07/2017
Accepted: 01/11/2018

Keywords

Average
Scrambled response model
Randomized response
Sensitive surveys
Privacy protection

Abstract

In this article, we use additive scrambling to estimate the mean of a sensitive variable. In the proposed scrambling model, taking G (>1) as a positive integer chosen by the interviewer, each respondent is asked to randomly draw *G* values from a given distribution of scrambling variable and add average of these randomly drawn values to his/her true response on the sensitive variable. Using repetition of the scrambling experiment, we propose a relatively more efficient estimator of sensitive mean without incurring any additional sampling cost. We present a generalization of additive scrambled response models and show that most of additive scrambling models are special cases of suggested generalization. Through algebraic and numerical comparisons, superiority of the proposed methodology is established.

1. INTRODUCTION

Asking sensitive questions directly is always a tricky issue because respondents may feel insecure and become reluctant to divulge the true response. Sensitive questions are related to habitual tax evasion, abortion, drunken driving, gambling, drug abuse, number of bottles of vine consumed in a month, number of violations of traffic rules in a year, number of cheatings in exam during the whole educational career, etc. Respondents mostly give evasive answers or refuse to give response on sensitive questions. [1] introduced a technique called Randomized Response Technique (RRT), consisting of two complementary questions; one of them is randomly chosen by the respondents and answered truthfully. The respondent only reports a 'yes' or 'no' and does not unveil the question randomly selected by him/her. Through this RRT, honest and reliable data are gathered and unbiased estimate of the proportion of individuals possessing sensitive attribute is obtained. This method proved very effective in medical, socioeconomic, biological and many other fields of life to collect trustworthy data.

After the pioneering work of [1], different contributions towards randomized response sampling have been put forward in the literature. [2] and [3,4] introduced an unrelated question technique consisting of two questions; a question about sensitive variable and a question about unrelated (non sensitive) variable. By using this technique, privacy of the respondent is more protected. [5-15] modified the work of [1] to improve the efficiency of the estimators and provided increased privacy protection. Some more work can be seen in the following papers of [16-21] etc.

There are three types of scrambled response models we found in the literature, namely, additive scrambled response models (ASRM), multiplicative scrambled response models (MSRM) and mixed scrambled response models. For each model to be discussed later, the reported response is denoted by Z. The corresponding estimators and their variances differ in the reported response Z, its sample and population means $\bar{Z}$ and μ_Z and its population variance σ_Z^2. The traditional ASRM was introduced by [22]. In this model, a scrambling variable (with known distribution) is added to the true sensitive response. This model

*Corresponding author, e-mail: zhlangah @yahoo.com

provides good privacy protection to the respondent and may be briefly explained as follows. Let X be a study variable with mean μ_X and variance σ_X^2. Let Y be a scrambling variable with mean $\mu_Y = 0$ and variance σ_Y^2. The respondent chosen by simple random sampling gives the additive scrambled response as

$$Z = X + Y. \tag{1}$$

The unbiased estimator, $\hat{\mu}_{XH}$, of population mean and its variance are given, respectively, by

$$\hat{\mu}_{XH} = \bar{Z}, \tag{2}$$

and

$$Var(\hat{\mu}_{XH}) = \frac{\sigma_X^2 + \sigma_Y^2}{n}. \tag{3}$$

The above model may also be termed as a zero stage model.

[23] proposed a two stage optional randomized response model by using a partial scrambling approach. Two subsamples are required in this approach. In each subsample, a proportion P of respondents give the true response and remaining proportion $(1-P)$ of the respondents is directed to go to the second stage.

Every respondent, using second stage, has the option to report on the sensitive question if he/she feels the question insensitive or to report an additive scrambled response, otherwise. Let W be the population proportion of individuals who feels the study question as sensitive. Let X_i be the study variable with mean μ_X and variance σ_X^2. Let Y_i be the scrambling variable with mean $\theta_i, i = 1,2$ and variance $\sigma_{Y_i}^2$. The total sample size n is divided into two subsamples of sizes n_1 and n_2. The distribution of the reported response in the i^{th} subsample is given by

$$Z_i = \begin{cases} X & with\ probability\ P + (1-P)(1-W) \\ X + Y_i & with\ probability \quad W(1-P) \end{cases} \tag{4}$$

An unbiased estimator of population mean and its variance are given, respectively, by

$$\hat{\mu}_{XG} = \frac{\theta_2 \overline{Z_1} - \theta_1 \overline{Z_2}}{\theta_2 - \theta_1} \tag{5}$$

and

$$Var(\hat{\mu}_{XG}) = \frac{1}{(\theta_2 - \theta_1)^2}\left\{\theta_2^2 \frac{\sigma_{Z1}^2}{n_1} + \theta_1^2 \frac{\sigma_{Z2}^2}{n_2}\right\}, \tag{6}$$

where $\sigma_{Z_i}^2 = \sigma_X^2 + \sigma_{Y_i}^2\{(1-P)W\} + \theta_i^2\{(1-P)W\}[1-(1-P)W]$.

[24] proposed a different two stage procedure. At 1st stage, a proportion T of the respondents gives the scrambled response and the remaining proportion $(1-T)$ of the respondents has two options; either to give true response with probability $(1-W)$ or to give additive scrambled response with probability W, where W is the sensitivity level. Let X be the study variable with mean μ_X and variance σ_X^2. Let Y_i be the scrambled variable with mean $\mu_{Y_i}, i = 1,2$ and variance $\sigma_{Y_i}^2$. The total sample size n is divided into two subsamples of sizes n_1 and n_2. The observed response Z_i from the i^{th} subgroup may be written as

$$Z_i = \begin{cases} X & with\ probability\ (1-T)(1-W) \\ X + Y_i & with\ probability\ T + W(1-T) \end{cases} \tag{7}$$

An unbiased estimator of population mean and its variance are given, respectively, by

$$\hat{\mu}_{XM} = \frac{\theta_2 \overline{Z_1} - \theta_1 \overline{Z_2}}{\theta_2 - \theta_1} \tag{8}$$

and

$$Var(\hat{\mu}_{XM}) = \frac{1}{(\theta_2 - \theta_1)^2}\left\{\theta_2^2 \frac{\sigma_{Z1}^2}{n_1} + \theta_1^2 \frac{\sigma_{Z2}^2}{n_2}\right\}, \tag{9}$$

where $\sigma_{Z_i}^2 = \sigma_X^2 + \sigma_{Y_i}^2\{T + (1-T)W\} + \theta_i^2\{T + (1-T)W\}[1 - [T + (1-T)W]$.

[21] proposed a scrambling model based on two random subsamples from the population. The respondents are requested to give the true response or to give a scrambled response. In this model, two scrambling variables are used. Let Y_i be the first scrambling variable with mean $\mu_{Y_i} = 1$ and variance $\sigma_{Y_i}^2$. Let X be the study variable with mean μ_X and variance σ_X^2. Let S_i be the second scrambling variable with mean $\mu_{S_i} = \theta_i$ and variance $\sigma_{S_i}^2 = \gamma_i^2$. The total sample size n is divided in two subsamples of sizes n_1 and n_2. The optional randomized response model for the i^{th} subsample is given by

$$Z_i = (1-T)X + T(S_i X + Y_i), \tag{10}$$

where $T \sim Bernoulli(W)$ and W is sensitivity level. The unbiased estimator of population mean and its variance are given, respectively, by

$$\hat{\mu}_{XHU} = \frac{\theta_1 \bar{Z}_2 - \theta_2 \bar{Z}_1}{\theta_1 - \theta_2} \tag{11}$$

and

$$Var(\hat{\mu}_{XHU}) = \frac{1}{(\theta_1 - \theta_2)^2}\left\{\theta_1^2 \frac{\sigma_{Z_2}^2}{n_2} + \theta_2^2 \frac{\sigma_{Z_1}^2}{n_1}\right\}, \tag{12}$$

where $\sigma_{Z_i}^2 = \sigma_X^2 + W\sigma_{Y_i}^2(\sigma_X^2 + \mu_X^2) + W\gamma_i^2 + W(1-W)\theta_i^2$.

[25] modified the [21] model and proposed a two stage model. At the 1st stage the respondents have to give the true response with probability P or to move to second stage with probability $1 - P$. The 2nd stage is the same as that of [21]. The response of the i^{th} respondent is given by

$$Z_i = \begin{cases} X & with\ probability\ P + (1-P)(1-W) \\ S_i X + Y_i & with\ probability \quad W(1-P) \end{cases} \tag{13}$$

An unbiased estimator of population mean and its variance are given, respectively, by

$$\hat{\mu}_{XG2} = \frac{\theta_2 \bar{Z}_1 - \theta_1 \bar{Z}_2}{\theta_2 - \theta_1} \tag{14}$$

and

$$Var(\hat{\mu}_{XG2}) = \frac{1}{(\theta_1 - \theta_2)^2}\left\{\theta_2^2 \frac{\sigma_{Z_1}^2}{n_1} + \theta_1^2 \frac{\sigma_{Z_2}^2}{n_2}\right\}, \tag{15}$$

where $\sigma_{Z_i}^2 = \sigma_X^2 + W(1-P)\sigma_{Y_i}^2(\sigma_X^2 + \mu_X^2) + W(1-P)\left(\gamma_i^2 + \theta_i^2\right) - W^2(1-P)^2\theta_i^2$.

MSRM was proposed by Eichhorn and Hayre (1983). In this model, the scrambling variable, having a known distribution, is multiplied by the true response. Their model increases the privacy protection. [7] gave a one stage MSRM where respondents have the option to report a true response or a scrambled response. If a respondent feels question as insensitive, he/she reports the true value of sensitive variable, otherwise, he/ she reports a scrambled response.

Ryu et al. (2006) introduced a two stage MSRM. At 1st stage, the respondents are given two statements; either they can choose the direct question with probability P or move to 2nd stage with probability $1 - P$. At the 2nd stage the respondents give answer to the direct question with known probability T or a scrambled response with probability $1 - T$. Let X be the study variable with mean μ_X and variance σ_X^2. Let Y be the scrambling variable with mean μ_Y and variance σ_Y^2. An unbiased estimator of population mean and its variance are given, respectively, by

$$\hat{\mu}_X = \bar{Z} \tag{16}$$

and

$$Var(\hat{\mu}_X) = \frac{1}{n}\{\sigma_X^2 + (1-P)T\sigma_Y^2(\sigma_X^2 + \mu_X^2)\} \tag{17}$$

[15] modified the Ryu et al. (2006) model by modifying the 2nd stage. At 2nd stage, respondents give the true response with probability $1 - W$ or give scrambled response with probability W, where W is termed as sensitivity level of the sensitive variable.

It is interesting to see that the reported response from all the ASRMs, discussed above, can be written in a generalized form as:

$$Z = \alpha_i X_i + (1 - \alpha_i)(X_i + Y_i), \tag{18}$$

where X is the study variable with mean μ_X and varianceσ_X^2, Y is the scrambling variable with mean μ_Y and variance σ_Y^2, and α_i is Bernoulli random variable with $E(\alpha_i) = A$. Then an unbiased estimator of population mean and its variance are given, respectively, by

$$\hat{\mu}_X = \bar{Z} - (1 - A)\mu_Y \tag{19}$$

and

$$Var(\hat{\mu}_X) = \frac{1}{n}\{\sigma_X^2 + (1-A)(\sigma_Y^2 + A\mu_Y^2)\}. \tag{20}$$

[28] proposed a new additive scrambled response model in which they use three statements. Each respondent is asked to rotate the spinner and report the answer to one of statements, randomly pointed by the spinner. The three statements are: (i) Report the true response X with probability, p_1, (ii) Report the

scrambled response $XS_1 + S_2$ with probability, p_2 and (iii) Report the value of S_3 with probability p_3. The variables S_1, S_2 and S_3 are three scrambling variables. If we replace $p_1 = A, p_2 = 1 - A, S_1 = 1, S_2 = Y$ and $S_3 = 0$ then the generalized model given in (18) becomes a special case of the [28] model.

The rest of the article is arranged as follows. In section 2, we present proposed generalized model and show that most of the existing ASRMs are special cases of the proposed model and a list of some new models is also given. In section 3, we compare proposed model with existing models discussed in Section 1, and prove the superiority of our proposed models both analytically and numerically. In section 4, we measure the privacy protection of respondent. In the last section, we give the summary and concluding remarks of this article.

2. PROPOSED MODEL

Let we have a population $U = (u_1, u_2, \dots, u_N)$ of size *N* and a random sample of size *n* is selected from the population. Respondents are selected by simple random sampling with replacement. Let *A* be the proportion of sampled individuals asked to answer truthfully and the remaining proportion, $(1 - A)$, of the respondents is directed to report a scrambled response. The scrambling is done as follows. Each respondent (who randomly chooses to scramble his/her response) is provided a randomization device to generate the $G(> 1)$ values of scrambling variable (following a pre-assigned distribution). For a given respondent, sample mean of scrambling variable is $\bar{y} = \frac{\sum_{j=1}^{G} Y_j}{G}$. Then he/she is requested to add average $\bar{y}$ of G values of scrambling variable to his/her true response. The randomization (scrambling) procedure is performed in such a way that the interviewer doesn't know about the G values of scrambling variable. Let X be the study variable with mean μ_X and variance σ_X^2. Let *Y* be the scrambling variable with mean μ_Y and variance σ_Y^2. Now, the i^{th} reported response Z_i is given by

$$Z_i = \begin{cases} X_i & with\ probability\ A \\ X_i + \bar{y}_i & with\ probability\ 1 - A \end{cases} \tag{21}$$

The expected response is given by

$$E(Z) = AE(X) + (1 - A)E(X + \bar{Y})$$

$$E(Z) = A\mu_X + (1 - A)(\mu_X + \mu_Y)$$

$$E(Z) = \mu_X + (1 - A)\mu_Y.$$

An estimator of the population mean μ_X may be suggested as:

$$\hat{\mu}_{XP} = \bar{Z} - (1 - A)\mu_Y. \tag{22}$$

The variance of the proposed estimator $\hat{\mu}_{XP}$ is given by

$$Var(\hat{\mu}_{XP}) = \frac{1}{n}\left\{\sigma_X^2 + (1 - A)\left(\frac{\sigma_Y^2}{G} + A\mu_Y^2\right)\right\}. \tag{23}$$

Theorem 1: $\hat{\mu}_{XP}$is an unbiased estimator of population mean μ_X.

Proof: Theorem 1 can easily be proved by using the fact that $E(\bar{Z}) = E(Z)$.

Theorem 2: The variance of the unbiased estimator $\hat{\mu}_{XP}$ is given by

$$Var(\hat{\mu}_{XP}) = \frac{1}{n}\left\{\sigma_X^2 + (1 - A)\left(\frac{\sigma_Y^2}{G} + A\mu_Y^2\right)\right\}.$$

Proof: We have

$$Var(\hat{\mu}_{XP}) = \frac{\sigma_Z^2}{n}.$$

To find σ_Z^2 , consider

$\sigma_Z^2 = E\left(Z_i^2\right) - \left(E(Z_i)\right)^2$.

and

$E\left(Z_i^2\right) = \sigma_X^2 + \mu_X^2 + (1-A)\left\{\frac{\sigma_Y^2+\mu_Y^2}{G} + \frac{G-1}{G}\mu_Y^2\right\} + 2(1-A)\mu_X\mu_Y$.

Hence, the Theorem 2 can easily be proved.

Theorem 3: The unbiased estimator of the variance of $\hat{\mu}_{XP}$ is given by

$\widehat{Var}(\hat{\mu}_{XP}) = \frac{s_Z^2}{n}$,

where $s_Z^2 = \frac{1}{n-1}\sum_i^n (Z_i - \bar{Z})^2$.

Proof: Theorem 3 can easily be proved by using the fact that $E(s_Z^2) = \sigma_Z^2$.

All the ASRM, discussed earlier, can be written as one stage model and, now, we show that all the models are special cases of the proposed model. Corresponding to [22] model, a new model may be proposed by setting $A = 0$ in Equation (21) as follows.

$$Z_i = X + \bar{y}_i \tag{24}$$

Now, the unbiased estimator of population mean and its variance are given, respectively, by

$$\hat{\mu}_{XP1} = \bar{Z} \tag{25}$$

and

$$Var(\hat{\mu}_{XP1}) = \frac{(\sigma_X^2+\sigma_Y^2/G)}{n} \quad . \tag{26}$$

Corresponding to the [23] model, distribution of reported response through proposed methodology is given by

$$Z_i = \begin{cases} X & with\ probability\ P + (1-P)(1-W) \\ X + \bar{Y} & with\ probability \quad W(1-P) \end{cases} \tag{27}$$

Using the above reported response, an unbiased estimator of population mean and its variance are now given, respectively, by

$$\hat{\mu}_{XP2} = \frac{\theta_2\overline{Z_1} - \theta_1\overline{Z_2}}{\theta_2 - \theta_1} \tag{28}$$

and

$$Var(\hat{\mu}_{XP2}) = \frac{1}{(\theta_2-\theta_1)^2}\left\{\theta_2^2\frac{\sigma_{Z1}^2}{n_1} + \theta_1^2\frac{\sigma_{Z2}^2}{n_2}\right\}, \tag{29}$$

where

$\sigma_{Zi}^2 = \sigma_X^2 + \frac{\sigma_Y^2}{G}\{(1-P)W\} + \theta_i^2\{(1-P)W\}[1-(1-P)W]$.

Corresponding to the [24] model, the proposed structure generates the following distribution of reported response.

$$Z_i = \begin{cases} X & with\ probability\ (1-T)(1-W) \\ X + \bar{Y} & with\ probability\ T + W(1-T) \end{cases} \tag{30}$$

An unbiased estimator of population mean and its variance are given, respectively, by

$$\hat{\mu}_{Xp3} = \frac{\theta_2\overline{Z_1} - \theta_1\overline{Z_2}}{\theta_2 - \theta_1} \tag{31}$$

and

$$Var(\hat{\mu}_{Xp3}) = \frac{1}{(\theta_2-\theta_1)^2}\left\{\theta_2^2\frac{\sigma_{Z1}^2}{n_1} + \theta_1^2\frac{\sigma_{Z2}^2}{n_2}\right\}, \tag{32}$$

where $\sigma_{Zi}^2 = \sigma_X^2 + \frac{\sigma_Y^2}{G}\{T + (1-T)W\} + \theta_i^2\{T + (1-T)W\}[1 - [T + (1-T)W]]$.

Following the proposal by [21], the expected response through the proposed model may be written as:

$$\mathrm{E(Z_i)} = \mathrm{(1\text{-}T)X + T(S_iX + \overline{Y})}. \tag{33}$$

Based on the above reported response, an unbiased estimator of population mean and its variance are given, respectively, by

$$\hat{\mu}_{XP4} = \frac{\theta_1\bar{Z}_2 - \theta_2\bar{Z}_1}{\theta_1-\theta_2} \tag{34}$$

and

$$Var(\hat{\mu}_{XP4}) = \frac{1}{(\theta_1-\theta_2)^2}\left\{\theta_1^2\frac{\sigma_{Z2}^2}{n_2} + \theta_2^2\frac{\sigma_{Z1}^2}{n_1}\right\}, \tag{35}$$

where $\sigma_{Zi}^2 = \sigma_X^2 + W\frac{\sigma_Y^2}{G}(\sigma_X^2 + \mu_X^2) + W\gamma_i^2 + W(1-W)\theta_i^2$.

Similarly, corresponding to the [25] model, the proposed structure would generate the distribution of reported response given by

$$Z_i = \begin{cases} X & with\ probability\ P + (1-P)(1-W) \\ S_iX + \bar{Y} & with\ probability \quad W(1-P). \end{cases} \tag{36}$$

And, an unbiased estimator of population mean and its variance are given, respectively, by

$$\hat{\mu}_{XP5} = \frac{\theta_2\bar{Z}_1 - \theta_1\bar{Z}_2}{\theta_2-\theta_1} \tag{37}$$

and

$$Var(\hat{\mu}_{XP5}) = \frac{1}{(\theta_1-\theta_2)^2}\left\{\theta_2^2\frac{\sigma_{Z1}^2}{n_1} + \theta_1^2\frac{\sigma_{Z2}^2}{n_2}\right\}, \tag{38}$$

where $\sigma_{Zi}^2 = \sigma_X^2 + W(1-P)\frac{\sigma_Y^2}{G}(\sigma_X^2 + \mu_X^2) + W(1-P)\left(\gamma_i^2 + \theta_i^2\right) - W^2(1-P)^2\theta_i^2$.

It is interesting to note that for $G = 1$, all the proposed models reduce to usual corresponding models.

If we replace $p_1 = A$, $p_2 = 1-A$, $S_1 = 1$, $S_2 = \bar{Y}$ and $S_3 = 0$ in [28] model then our proposed generalized model becomes a special case of the [28] model.

3. EFFICIENCY COMPARISON

In this section, we carry out some analytical and numerical comparisons between the proposed models and some existing models.

Consider,

$Var(\hat{\mu}_{XP}) < Var(\hat{\mu}_X)$

Using Equations (20) and (23) in the above inequality, we get

$$\frac{1}{n}\left\{\sigma_X^2 + (1-A)(\frac{\sigma_Y^2}{G} + A\mu_Y^2)\right\} < \frac{1}{n}\{\sigma_X^2 + (1-A)(\sigma_Y^2 + A\mu_Y^2)\}$$

$$(1-A)(\frac{\sigma_Y^2}{G}+A\mu_Y^2) < (1-A)(\sigma_Y^2+A\mu_Y^2)$$

$$\frac{\sigma_Y^2}{G} < \sigma_Y^2$$

$$G > 1\ ,$$

which is always true since in the proposed structure we assumed $G > 1$. Thus, the proposed methodology is better than the usual one where a single value of scrambling variable is used to scramble the response.

To know the extent of relative efficiency and prove the superiority of the proposed model a simulation study is conducted. The simulated variances for proposed and existing models (when $G = 1$) are calculated using random samples of size 100. Different values of *A* such as 0.1, 0.3, 0.5,0.7 and 0.9 are considered.

The number of values of scrambling variable are considered as $G = 1, 3$ and 5. The study variable *X* is assumed to be Normally distributed with mean 2 and variance 1. The scrambling variable *Y* is taken as Poisson random variable with mean $\mu_Y = 3$. The simulated results based on 50000 iterations are showcased in Table 1, given below.

Table 1. *Comparison Table of* $Var(\hat{\mu}_X)$ *and* $Var(\hat{\mu}_{XP})$

G	*A*				
	0.1	0.3	0.5	0.7	0.9
1	0.04492325	0.04987604	0.0470403	0.03810669	0.02096469
3	0.02711408	0.03569062	0.03749994	0.03204375	0.01884359
5	0.02400066	0.03326997	0.03516758	0.03066921	0.01850494

The results in the Table 1 clearly show that the variance of the estimator based on the proposed generalized model is less than the variances of the estimator when $G = 1$. The superiority of the proposed model can be easily proved by taking any of the models discussed in this paper and by taking any distribution of scrambling variable.

4. PRIVACY PROTECTION

In this section, we calculate the respondent's degree of protection. That is, we try to find how much secure respondents feel when using the proposed model. We find the loss in privacy protection after using proposed model. A good statistical analysis or planning in randomized response is that through which we increase both the efficiency and privacy protection. We use a criterion proposed by [29]. To find degree of privacy protection, denoted by Δ, we take the expected value of the square of the difference of the response obtained by randomized response technique and the true response. Let *R* be the response of the respondent obtained by using any randomized response technique and *X* be the true response, then the measure of privacy protection defined by [29] is given by

$$\Delta = E(R-X)^2\ .$$

We use generalized model and proposed generalized model defined in Section 1 and 2 to check loss in privacy protection. The amount of privacy protection, Δ_X, for the generalized model is

$$\Delta_X = E(R-X)^2$$

$$\Delta_X = E(\alpha_i X_i + (1-\alpha_i)(X_i+Y_i) - X)^2$$

$$\Delta_X = (1-A)(\mu_Y^2 + \sigma_Y^2)$$

Similarly, the amount of privacy protection, Δ_{XP}, for the proposed generalized model, explained in Section 2, is

$$\Delta_{XP} = E(R - X)^2$$

$$\Delta_{XP} = E(AX + (1 - A)(X + \bar{Y}) - X)^2$$

$$\Delta_{XP} = (1 - A)(\mu_Y^2 + \frac{\sigma_Y^2}{G})$$

Now, the loss in amount of privacy protection is

$$\Delta_X - \Delta_{XP} = (1 - A)(\mu_Y^2 + \sigma_Y^2) - (1 - A)(\mu_Y^2 + \frac{\sigma_Y^2}{G})$$

$$\Delta_X - \Delta_{XP} = (1 - A)\sigma_Y^2 \left(\frac{G-1}{G}\right). \quad (39)$$

To have a tangible idea about the loss in privacy, we calculate the percentage relative loss (PRL) defined as $\left(\frac{\Delta_X - \Delta_{XP}}{\Delta_X}\right) \times 100$ for $G = 2, 3$, and different values of μ_Y. The results for PRL are presented in the Table 2.

Table 2. *The PRL for different values of* μ_Y *and* σ_Y^2

	$G=2$					$G=3$				
	$\sigma_Y^2 = 2.5$	$\sigma_Y^2 = 3$	$\sigma_Y^2 = 3.5$	$\sigma_Y^2 = 4$	$\sigma_Y^2 = 4.5$	$\sigma_Y^2 = 2.5$	$\sigma_Y^2 = 3$	$\sigma_Y^2 = 3.5$	$\sigma_Y^2 = 4$	$\sigma_Y^2 = 4.5$
$\mu_Y = 2$	27.77	30.00	31.81	33.33	34.61	37.03	40.00	42.42	44.44	46.15
$\mu_Y = 2.5$	25.00	27.27	29.16	30.30	32.14	33.33	36.36	38.88	41.02	42.85
$\mu_Y = 3$	22.72	25.00	26.92	28.57	30.00	30.30	33.33	35.89	38.09	40.00
$\mu_Y = 3.5$	20.83	23.07	25.00	26.60	28.12	27.77	30.76	33.33	35.55	37.00
$\mu_Y = 4$	19.23	21.42	23.37	25.00	26.47	25.64	28.57	31.11	33.33	35.29

From (39), we can see that as the value of *G* increases, the loss in privacy protection increases and for $G = 1$, both the models provide equal privacy protection. Also, from Table 2, it is observed that PRL decreases (increases) with the increase in $\mu_Y \left(\sigma_Y^2\right)$. As is obvious, for G>1, our proposed model performs better than the other models in terms of efficiency. So, we check how much we gain in efficiency.

$$Gain\ in\ efficiency = Var(\hat{\mu}_X) - Var(\hat{\mu}_{XP})$$

$$= \frac{1}{n}\{\sigma_X^2 + (1 - A)(\sigma_Y^2 + A\mu_Y^2)\} - \frac{1}{n}\left\{\sigma_X^2 + (1 - A)(\frac{\sigma_Y^2}{G} + A\mu_Y^2)\right\}$$

$$= \frac{(1-A)}{n}(\sigma_Y^2 + A\mu_Y^2) - \frac{(1-A)}{n}\left(\frac{\sigma_Y^2}{G} + A\mu_Y^2\right)$$

$$=\frac{(1-A)}{n}\sigma_Y^2\left(\frac{G-1}{G}\right). \quad (40)$$

Again, to have a clear picture of gain in efficiency, we calculate the percentage relative gain (PRG) defined as $\left(\frac{Var(\hat{\mu}_X)-Var(\hat{\mu}_{XP})}{Var(\hat{\mu}_X)}\right)\times 100$ for $G=2,3$, and different values of σ_X^2. The numerical values of PRG so obtained are arranged in the Table 3.

Table 3. *The PRG for* σ_X^2 *and A*

	$G=2$					$G=3$				
	$A=0.1$	$A=0.3$	$A=0.5$	$A=0.7$	$A=0.9$	$A=0.1$	$A=0.3$	$A=0.5$	$A=0.7$	$A=0.9$
$\sigma_X^2=1$	32.80	30.12	27.60	25.16	22.73	43.73	40.16	36.80	33.55	30.30
$\sigma_X^2=1.5$	28.62	26.17	23.84	21.55	19.23	38.16	34.90	31.78	28.73	25.64
$\sigma_X^2=2$	25.40	23.15	20.99	18.84	16.70	33.86	30.86	27.97	25.12	22.22
$\sigma_X^2=2.5$	22.82	20.74	18.73	16.74	14.70	30.42	27.66	24.95	22.32	19.60
$\sigma_X^2=2$	20.71	18.79	16.92	15.00	13.15	27.62	25.06	22.56	20.08	17.54

From the results in Table 3, it is observed that PRG can be increased by increasing the value of G. Similarly, the PRG is higher for smaller values of A and σ_X^2.

From (39) and (40), it is clearly seen that we have loss in privacy protection and gain in efficiency when $G>1$is used. [29] proved that in every randomized response the efficiency decreases when amount of privacy protection increases and vice versa. They showed that increase in privacy and increase in efficiency are in conflict. If, we use the direct questioning method then we have the minimum variance ($i.e.$ $\frac{\sigma_X^2}{n}$) but, in this case, we have to sacrifice the privacy of the respondent (amount of privacy protection is zero). In this article, we see that as the value of *G* increases we have smaller variance of the proposed estimator.

According to central limit theorem, as the value of G increases, the distribution of $\bar{Y}$ approaches to normal distribution with mean μ_Y. As the value of *G* increases, we know that value of $\bar{Y}$ would be closer to μ_Y with higher probability. Thus, we would actually have good guess about the true response of a respondent. So, we definitely lose privacy of respondent for larger value of *G*. Intuitively, we suggest that *G* should be less than 5 since respondents are less burdened as they are required to select a few values of scrambling variable. We also found this suggestion of fixing the number of values of scrambling variable at 2, 3 and 4 in [30]. Our intuition may also be supported by the observations from the Table 4, given below. From Table 4, it is obvious that most (75%) of the loss and gain is achieved at $G=4$. Moving from $G=4$to $G=10$ results in only 15% increase in "loss in privacy" and "gain in efficiency". Obviously, when we move from $G=4$ to onwards, only 25% increase in "loss in privacy" and "gain in efficiency" will be observed. Thus, to keep the proposed model more parsimonious setting $G<5$ seems more practicable.

Table 4. *Percentage of "loss in privacy" and "gain in efficiency" for different values of G*

G	2	3	4	5	6	7	8	9	10
Loss and Gain	50%	67.66%	75%	80%	83.33%	85.71%	87.5%	88.88%	90%

5. SUMMARY AND CONCLUSION

The idea of increasing the protection of respondent's privacy through scrambling of the response is very common today. The researchers use randomized response devices to perturb the value of sensitive variable by adding, subtracting or multiplying the value of scrambling variable whose distribution is known. In this article, we gave the respondent a new device which requires adding average of *G* random values of a scrambling variable to the value of sensitive variable. The main advantage behind this idea is that it increases the respondent's cooperation without increasing the sampling cost and loosing efficiency. As special cases, different additive scrambled response models are discussed. A complete theoretical comparison of existing models and the proposed model is presented and superiority of the proposed model is established. A simulation based study is also conducted to know the extent of better performance of the proposed method. We also checked the respondent privacy protection whether it remains same or not. We conclude that for $G>1$, we have to sacrifice same percentage of amount of respondent privacy as the increase in efficiency.

CONFLICTS OF INTEREST

No conflict of interest was declared by the authors.

REFERENCES

[1] Warner, S. L., "Randomized response: a survey technique for eliminating evasive answer bias", Journal of the American Statistical Association, 60:63-69, (1965).

[2] Horvitz, D. G., Shah, B. V., Simmons, W. R., "The unrelated question randomized response model", Proceedings of Socialogical Statistics Section of American Statistical Association, 65-72, (1967).

[3] Greenberg, B. G., Abul-Ela Abdel-Latif, A., Simmons, W. R., Horvitz, D. G., "The unrelated question rr model: theoretical framework", Journal of the American Statistical Association, 64: 52-539, (1969).

[4] Greenberg, B.O., Kuebler, R. R. Jr., Abernathy, J. R., Horvitz, D. G., Application of the randomized response technique in obtaining quantitative data", Journal of the American Statistical Association, 66:243-250, (1971).

[5] Mangat, N. S., "An ımproved randomized response strategy", Journal of the Royal Statistical Society Series B, 56: 93-95, (1994).

[6] Singh, S., Joarder, A. H., "Unknown repeated trials in randomized response sampling", Journal of Indian Society of Agricultural Statistics, 50(1): 103-105, (1997).

[7] Gupta, S. N., Gupta, B. C., Singh, S., Estimation of sensitivity level of personal ınterview survey questions", Journal of Statistical Planning and Inference, 100:239–247, (2002).

[8] Arnab, R., "Optional randomized response techniques for complex survey designs", Biometrical Journal, 46:114–124, (2004).

[9] Mangat, N. S., Singh, R., Singh, S., "Violation of respondent's privacy in moor's model—ıts rectification through a random group strategy", Communication in Statistics-Theory and Methods, 26:743–754, (1997).

[10] Singh, S., Singh, R., Mangat. N. S., "Some alternative strategies to moor's model in randomized response sampling- a survey technique for eliminating evasive answer bias", Journal of Statistical Planning and Inference, 83: 243-255, (2002).

[11] Chang, H. J., Huang, K. C., "Estimation of proportion and sensitivity of a qualitative character", Metrika, 53:269-280, (2001).

[12] Bhargava, M., Singh, R., "A modified randomization device for warner's model", Statistica, 60:315-321, (2000).

[13] Singh, S., Horn, S., Singh, R., Mangat, N. S., "On the use of modified randomization device for estimating the prevalence of a sensitive attribute", Statistics in Transition, 6(4): 515-522, (2003).

[14] Chang, H. J., Wang, C. L., Huang, K. C., "Using randomized response to estimate the proportion and truthful reporting probability in a dichotomous finite population", Journal of Applied Statistics, 31:565-573, (2004).

[15] Gupta, S. N., Thornton, B., Shabbir, J., Singhal, S, "A comparison of multiplicative and additive optional rrt models", Journal of Statistical Theory and Application, 5:226-239, (2006).

[16] Kim, J.M., Elam, M.E., "A Stratified Unrelated Randomized Response Model", Statistical Papers, 48: 215–233, (2007).

[17] Chaudhuri, A., Pal, S., "Estimating sensitive proportions from warner's randomized responses in alternative ways restricting to only distinct units sampled", Metrika, 68:147–156, (2008).

[18] Huang, K. C., "Estimation for the sensitive characteristic using optional randomized response technique", Quality and Quantity, 42:679-686, (2008).

[19] Pal, S., "Unbiasedly estimating the total of a stigmatizing variable from complex survey on permitting options for direct or randomized responses", Statistical Papers, 49: 157–164, (2008).

[20] Diana, G. Perri, P. F., "Estimating a sensitive proportion through randomized response procedures based on auxiliary ınformation", Statistical Papers, 50(3): 661–672, (2009).

[21] Huang, K. C., "Unbiased estimators of mean, variance and sensitivity level for quantitative characteristics in finite population sampling", Metrika, 71:341–352, (2010).

[22] Himmelfarb, S., Edgell, S. E., "Additive constant model: a randomized response technique for eliminating evasiveness to quantitative response questions", Psycholigical Bulletin, 87:525–530, (1980).

[23] Gupta, S., Shabbir, J., Sehra, S., "Mean and Sensitivity Estimation in Optional Randomized Response Models", Journal of Statistical Planning and Inference, 140(10): 2870-2874, (2010).

[24] Mehta, S., Dass, B. K., Shabbir, J. and Gupta, S., "A three stage optional randomized response model, Journal of Statistical Theory and Practice, 6 (3): 417-427, (2012).

[25] Gupta, S., Mehta, S., Shabbir, J., Dass, B. K., "Generalized scrambling in quantitative optional randomized response models", Communications in Statistics - Theory and Methods, 42, 4034-4042, (2012).

[26] Eichhron, B. H., Hayre, L. S., "Scrambled randomized response methods for obtaining sensitive quantitative data, Journal of Statistical Planning and Inference, 7:307–316, (1983).

[27] Ryu, J. B., Kim, J. M., Heo, T. Y., Park, C. G., "On stratified randomized response sampling", Model Assisted Statistics and Application,1:31–36, (2006).

[28] Arcos, A., Rueda, M, D, M., Singh, S., "A generalized approach to randomised response for quantitative variables", Quality & Quantity, 49:1239–1256, (2015).

[29] Zaizai, Y., Wang, j., Lai, J., "An efficiency and protection degree-based comparison among the quantitative randomized response strategies, Communications in Statistics - Theory and Methods, 38:3, 400-408, (2008).

[30] Hussain, Z., "On eliminating the scrambling variance in scrambled response models, International Journal of Academic Research in Business and Social Sciences, 2 (6), 39-45, (2012).