Journal of Statistical Theory and Practice (2021) 15:3
https://doi.org/10.1007/s42519-020-00135-2

ORIGINAL ARTICLE

Mean Estimation of Sensitive Variables Under Non-response and Measurement Errors Using Optional RRT Models

Qi Zhang[1] · Sadia Khalil[2] · Sat Gupta[1]

Accepted: 2 October 2020 / Published online: 3 November 2020

Abstract
This study focuses on three issues we face in survey sampling: non-response, measurement errors, and social desirability bias. We propose a generalized mean estimator in the presence of measurement errors and non-response using optional RRT methodology under simple random sampling. We present a comparison of the proposed estimator with some commonly used estimators.

Keywords Non-response · Measurement errors · Optional RRT models · Mean square error

1 Introduction

Nowadays, many researchers use email or phone surveys which is an easier, cheaper, and faster way to obtain information. However, it causes a high non-response rate. This reduces the accuracy of parameter estimates. Among all the sampling methods, face-to-face interview is one that reduces non-response rate the most, but the cost is considerably higher than other methods. Hansen and Hurwitz [11] were the first to suggest a procedure of taking a subsample of non-respondents after the first email or phone attempt and then obtaining information from this group by personal interview.

✉ Qi Zhang
qizhangpersonal@gmail.com

Sadia Khalil
sadia_khalil@hotmail.com

Sat Gupta
sngupta@uncg.edu

1 Department of Mathematics and Statistics, University of North Carolina at Greensboro, Greensboro, NC 27402, USA

2 Department of Statistics, Lahore College for Women University, Lahore 54000, Pakistan

The problem of non-response has been discussed in many papers. Many researchers suggested different types of estimators for population parameters based on Hansen and Hurwitz [11] double sampling plan. Another method to increase the accuracy of population estimates is by using auxiliary information. Studies of mean estimation using information on auxiliary variables include Khare and Srivastava [14], Rao [22], Khare and Sinha [15–17], Kumar and Singh [19], Yaqub et al. [28], Bhushan and Pandey [3, 4], and Unal and Kadilar [26].

Hansen and Hurwitz [11] method could obtain more information from face-to-face interview in the second phase, but it may also cause non-response bias if the variable of interest is sensitive in nature. The respondents are unlikely to provide true response in face-to-face interview for such questions. To reduce the social desirability bias (SDB) caused by sensitive questions, one could use randomized response technique (RRT) models when we target the group of non-respondents. Subjects may refuse to respond on the first call but may provide scrambled response on the second call with personal interview. Diana et al. [6] proposed an unbiased population mean estimator under this two-phase sampling. Their estimator reduces non-response but increases the estimator variance due to the use of RRT model in the non-respondent group. Later, Ahmed et al. [1] proposed generalized ratio and regression estimators utilizing known coefficient of variation of the study variable in case of second sample by using RRT approach. This estimator improved efficiency when the auxiliary variable and study variable are highly correlated. Makhdum et al. [21] also proposed a generalized class of estimators for a sensitive study variable in the presence of non-response using RRT model.

Measurement error is another important issue in sample surveys. Most of the time we assume measurement errors to be very small and neglect them. But if measurement errors are not small enough, then we get unreliable estimates. Some existing studies which have discussed measurement errors in estimating population parameters include Kumar et al. [18], Kumar et al. [20], Khalil et al. [12], and Singh et al. [24]. Singh and Sharma [23], Singh and Vishwakarma [25], and Audu et al. [2] considered the problem of estimating the finite population mean in the presence of non-response and measurement errors simultaneously. Also, Khalil et al. [13] studied mean estimation under measurement errors using optional RRT models.

Based on the previous studies, one may consider estimating population mean of a sensitive variable in the simultaneous presence of both measurement error and non-response. This problem has not drawn much attention in the existing literature. RRT models used in the previous studies [1, 6, 21] are non-optional RRT models where all the respondents are required to provide a scrambled response. However, a survey question may be sensitive for one person but not for another. Gupta et al. [7] pointed out that if we give respondents the option to choose whether they want to answer the sensitive question directly or provide a scrambled response, the model would be more efficient while there is no extra loss of privacy [10].

We will briefly discuss the Hansen and Hurwitz [11] (HH) two-phase sampling procedure in Sect. 2.1 and the optional RRT (ORRT) model in Sect. 2.2. Some existing mean estimators are presented in Sect. 3.1, and a generalized mean estimator is introduced in Sect. 3.2. Section 4 provides the results of a simulation study, and Sect. 5 provides some concluding remarks.

2 Modified Hansen and Hurwitz [11] Procedure (HH)

2.1 Hansen and Hurwitz [11]: Two-Phase Sampling

Let $U = \{U_1, U_2, \dots, U_N\}$ be a finite population of size N and a random sample without replacement of size n is taken. We assume that only n_1 units provide response on the first call and remaining $n_2 = n - n_1$ units do not respond. Then, a subsample of size $n_s = \frac{n_2}{f}$ $(f > 1)$ is taken from the n_2 non-responding units. Hansen and Hurwitz [11] used mail survey at the first attempt and then used face-to-face interview at the second call.

Let $\mu_y = \frac{\sum_{i=1}^{N} y_i}{N}$ and $\sigma_y^2 = \frac{\sum_{i=1}^{N}(y_i-\mu_y)^2}{N-1}$, respectively, be the population mean and variance of the study variable y. Let $\mu_{y_{(1)}} = \frac{\sum_{i=1}^{N_1} y_i}{N_1}$ and $\sigma^2_{y_{(1)}} = \frac{\sum_{i=1}^{N_1}(y_i-\mu_{y_1})^2}{N_1-1}$, respectively, be the mean and variance of respondent group of size N_1, and $\mu_{y_{(2)}} = \frac{\sum_{i=1}^{N_2} y_i}{N_2}$ and $\sigma^2_{y_{(2)}} = \frac{\sum_{i=1}^{N_2}(y_i-\mu_{y_2})^2}{N_2-1}$, respectively, be the mean and variance of non-respondent group of size N_2. Then, the population mean is given by

$$\mu_y = W_1\mu_{y_{(1)}} + W_2\mu_{y_{(2)}}. \tag{1}$$

where $W_1 = \frac{N_1}{N}$ and $W_2 = \frac{N_2}{N}$. Not knowing N_1 poses a challenge of its own.

Let $\bar{y}_1 = \frac{\sum_{i=1}^{n_1} y_i}{n_1}$ be the sample mean for the response group, and $\bar{y}_2 = \frac{\sum_{i=1}^{n_s} y_i}{n_s}$ be the sample mean for non-response group. One may note here that $\bar{y}_1$ and $\bar{y}_2$ are unbiased estimators for $\mu_{y_{(1)}}$ and $\mu_{y_{(2)}}$, respectively.

Hansen and Hurwitz [11] suggested an unbiased population mean estimator given by

$$\bar{y} = w_1\bar{y}_1 + w_2\bar{y}_2, \tag{2}$$

where $w_1 = \frac{n_1}{n}$ and $w_2 = \frac{n_2}{n}$.

The variance of $\bar{y}$ is given by

$$\mathrm{Var}(\bar{y}) = \left(\frac{N-n}{Nn}\right)\sigma_y^2 + \frac{W_2(f-1)}{n}\sigma^2_{y_{(2)}}. \tag{3}$$

2.2 Optional RRT (ORRT) Models

Let Y be a sensitive study variable, and $y_i\,(i = 1, 2 \dots n)$ be a simple random sample without replacement from y_i $(i = 1, 2 \dots N)$. Let $\mu_y = \frac{1}{N}\sum_{i=1}^{N} y_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu_y)^2$, and $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$. Let T and S be the two scrambling variables with respective means μ_T and μ_S, and known variances σ_T^2 and σ_S^2. Let T, S, X and Y be mutually independent. The respondent is asked to report a scrambled response for the study variable (Y) if he/she considers the question sensitive, and a correct response otherwise.

One could use a simple additive RRT model where the scrambled response is given by $Y + S$ (as in Gupta et al. [8]), or one may use a more general RRT model where the scrambled response is given by $TY + S$ (as in Diana and Perri [5]). Note that the simple additive model is a special case of the second model if we let $\mathrm{Var}(T) = 0$ and $E(T) = 1$. Khaili et al. [13] showed that the simple additive model is more efficient but the general model has greater privacy. However, the general RRT model is better when we use a combined measure of efficiency and privacy $\delta = \frac{\mathrm{Var}(Z)}{\Delta}$ proposed by Gupta et al. [10], where Z is the scrambled response and $\Delta = E(Z - Y)^2$ is the privacy level for the same model, as given by Yan et al. [27]. One may note that the model with smaller δ value is preferred because it means either a larger privacy level, or smaller value of $\mathrm{Var}(\hat{\mu})$, or both. It may be observed that

$$\delta_{\text{additive RRT}} = 1 + \frac{\sigma_y^2}{\sigma_s^2} > 1 + \frac{\sigma_y^2}{\sigma_s^2 + \sigma_T^2(\mu_y^2 + \sigma_y^2)} = \delta_{\text{general RRT}}. \tag{4}$$

Hence, while working with the general RRT model, the scrambling variable T will put a burden on the model efficiency but will improve the privacy level. Overall, the general model is better in terms of the unified measure of efficiency and privacy.

Therefore, we will use the general scrambling model in this study. The optional version of model $Z = TY + S$ is given by

$$Z = \begin{cases} Y & \text{with probability } 1 - W \\ TY + S & \text{with probability } W, \end{cases} \tag{5}$$

where W is the probability that a respondent finds the question sensitive. The mean and variance, respectively, for Z are given by

$$E(Z) = E(Y)(1 - W) + E(TY + S)W = E(Y) \tag{6}$$

and

$$\mathrm{Var}(Z) = E(Z^2) - E^2(Z) = \sigma_y^2 + \sigma_S^2 W + \sigma_T^2(\sigma_y^2 + \mu_y^2)W. \tag{7}$$

Obviously optional RRT model is more efficient than the non-optional RRT model since variance of Z increases as W increases. When $W = 1$, the RRT model becomes a non-optional model.

2.3 Modified Version of Hansen and Hurwitz [11]: Two-Phase Sampling

In order to encourage the respondents to answer a sensitive survey question truthfully, we give the respondents the opportunity to scramble the response using ORRT in the second phase of HH procedure when there is a face-to-face interview. In this case, we are modifying the HH procedure assuming that in the first phase, the respondent group gives direct answer, and then in the second phase, ORRT model is used to get response from a subgroup of non-respondents.

From Sect. 2.2, we can write the general RRT model as $Z = (YT + S)J + Y(1 - J)$, where $J \sim$ Bernoulli(W). Therefore, $E(J) = W$, $\mathrm{Var}(J) = W(1 - W)$ and $E(J^2) = \mathrm{Var}(J) + E^2(J) = W$.

The expectation under randomization mechanism is given by

$$\begin{aligned} E_R(Z) &= E_R(TYJ + SJ + Y - YJ) \\ &= YE_R(TJ) + E_R(SJ) + Y - YE_R(J) \\ &= Y\mu_T W + \mu_S W + Y - YW \\ &= (\mu_T W + 1 - W)Y + \mu_S W. \end{aligned} \tag{8}$$

Also

$$\begin{aligned} V_R(Z) &= V_R(TYJ + SJ + Y - YJ) \\ &= V_R(TYJ) + V_R(SJ) + V_R(YJ) + 2\mathrm{Cov}(TYJ, SJ) - 2\mathrm{Cov}(TYJ, YJ) \\ &\quad - 2\mathrm{Cov}(SJ, YJ) \\ &= Y^2[(\sigma_T^2 + \mu_T^2)W - \mu_T^2 W^2] + [(\sigma_s^2 + \mu_S^2)W - \mu_S^2 W^2] + Y^2[W(1 - W) \\ &\quad + 2Y\mu_T\mu_S W(1 - W) - 2Y^2[\mu_T W(1 - W)] - 2Y[\mu_S W(1 - W)] \\ &= (Y^2\sigma_T^2 + \sigma_s^2)W. \end{aligned} \tag{9}$$

Let $\hat{y}_i$ be a transformation of the randomized response on the ith unit whose expectation under the randomization mechanism is the true response y_i. It is given by

$$\hat{y}_i = \frac{z_i - \mu_S W}{\mu_T W + 1 - W} \tag{10}$$

with

$$E_R(\hat{y}_i) = y_i \tag{11}$$

(from (8)), and

$$\begin{aligned} V_R(\hat{y}_i) &= \frac{V_R(z_i)}{(\mu_T W + 1 - W)^2} \\ &= \frac{[y_i^2\sigma_T^2 + \sigma_s^2]W}{(\mu_T W + 1 - W)^2} = \tau_i \end{aligned} \tag{12}$$

(from (9)).

With ORRT model, a modified version of the HH estimator is given by

$$\hat{\hat{y}} = w_1\bar{y}_1 + w_2\hat{\bar{y}}_2, \tag{13}$$

where $\hat{\bar{y}}_2 = \sum_{i=1}^{n_s}(\frac{\hat{y}_i}{n_s})$.

Let E_i and V_i be the expectation and variance in the ith phase ($i = 1, 2$) under the two-phase sampling. It is easy to verify that

$$
\begin{aligned}
E(\hat{\bar{y}}) &= E_1E_2[w_1\bar{y}_1 + w_2\hat{\bar{y}}_2] \\
&= E_1[w_1\bar{y}_1 + w_2E_R(\hat{\bar{y}}_2)] \\
&= E_1[w_1\bar{y}_1 + w_2\bar{y}_2)] \\
&= W_1\mu_{y_{(1)}} + W_2\mu_{y_{(2)}} \\
&= \mu_y
\end{aligned}
\tag{14}
$$

since $E_R(\hat{\bar{y}}_2) = \frac{1}{n_s}\sum_{i=1}^{n_s} E_R(\hat{y}_i) = \bar{y}_2$.

The variance of $\hat{\bar{y}}$ can be written as

$$
\begin{aligned}
\mathrm{Var}(\hat{\bar{y}}) &= E_1[V_2(\hat{\bar{y}})] + V_1[E_2(\hat{\bar{y}})] \\
&= E_1[V_2(w_1\bar{y}_1 + w_2\hat{\bar{y}}_2)] + V_1[E_2(w_1\bar{y}_1 + w_2\hat{\bar{y}}_2)] \\
&= E_1[0 + V_2(w_2\hat{\bar{y}}_2)] + V_1[w_1\bar{y}_1 + w_2\bar{y}_2] \\
&= E_1[V_2(w_2\hat{\bar{y}}_2)] + V_1(\bar{y}) \\
&= E_1[\frac{w_2^2}{n_s}\frac{\sum_{i=1}^{N_2}\frac{(y_i^2\sigma_T^2+\sigma_s^2)W}{(\mu_T W+1-W)^2}}{N_2}] + V(\bar{y}) \\
&= \mathrm{Var}(\bar{y}) + \frac{W_2 f}{n}\frac{\sum_{i=1}^{N_2}\tau_i}{N_2}.
\end{aligned}
\tag{15}
$$

Note $E(y_i^2) = \sigma_y^2 + \mu_y^2$, and

$$
E\left(\frac{w_2^2}{n_s}\right) = E\left(\frac{n_2^2}{n^2}\frac{f}{n_2}\right) = E\left(\frac{n_2 f}{n^2}\right) = \frac{f}{n^2}E(n_2) = \frac{f}{n^2}(nW_2) = \frac{W_2 f}{n}, \tag{16}
$$

if we assume $\frac{n}{N} \approx \frac{n_2}{N_2}$.

Since $\bar{y}$ is the original HH mean estimator, the variance of $\hat{\bar{y}}$ is given by

$$
\mathrm{Var}(\hat{\bar{y}}) = \theta\sigma_y^2 + \lambda\sigma_{y_{(2)}}^2 + \frac{W_2 f}{n}\left[\frac{[(\sigma_{y_{(2)}}^2 + \mu_{y_{(2)}}^2)\sigma_T^2 + \sigma_S^2]W}{(\mu_T W + 1 - W)^2}\right], \tag{17}
$$

where $\theta = \frac{(N-n)}{Nn}$ and $\lambda = \frac{(f-1)W_2}{n}$.

3 Mean Estimators Under Measurement Errors and Non-response

3.1 Existing Mean Estimators

Using the standard terminology, as used in Sect. 2.1, let $\mu_x = \frac{\sum_{i=1}^{N} x_i}{N}$ and $\sigma_x^2 = \frac{\sum_{i=1}^{N}(x_i-\mu_x)^2}{N-1}$, respectively, be the known population mean and variance of the auxiliary variable X. Let $\mu_{x_{(1)}} = \frac{\sum_{i=1}^{N_1} x_i}{N_1}$ and $\sigma_{x_{(1)}}^2 = \frac{\sum_{i=1}^{N_1}(x_i-\mu_{x_1})^2}{N_1-1}$, respectively, be the population mean and variance of the respondent group of size N_1, $\mu_{x_{(2)}} = \frac{\sum_{i=1}^{N_2} x_i}{N_2}$

and $\sigma^2_{x_{(2)}} = \frac{\sum_{i=1}^{N_2}(x_i-\mu_{x_2})^2}{N_2-1}$, respectively, be the population mean and variance of the non-respondent group of size N_2. Let $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$ be the correlation coefficient between X and Y. Similarly let $\rho_{xy_{(1)}} = \frac{\sigma_{xy(1)}}{\sigma_x\sigma_y}$ and $\rho_{xy_{(2)}} = \frac{\sigma_{xy(2)}}{\sigma_x\sigma_y}$, respectively, be the correlation coefficients between X and Y for the respondents group and the non-respondents group. Let the measurement error (ME) for the auxiliary variable (X) in the population be given by $V_i = x_i - X_i$. Let the respective ME associated with the study variable (Y) in the population and the scrambled variable (Z) in the face-to-face interview phase be given by $U_i = y_i - Y_i$ and $P_i = z_i - Z_i$. These measurement errors are assumed to be random and uncorrelated with mean zero and variances σ^2_v, σ^2_u, and σ^2_p, respectively.

Assume population mean μ_x of auxiliary variable is known, and non-response happened on both X and Y. ORRT version of some of the existing mean estimators are listed below.

1. An ordinary mean estimator for sensitive variable in a finite population under modified HH is given by

$$\hat{\mu}_{yw} = \hat{\bar{y}}^* = w_1\bar{y}_1 + w_2\bar{y}_2^*, \quad \text{where } \bar{y}_2^* = \frac{1}{n_s}\sum_{i=1}^{n_s} z_i. \tag{18}$$

The MSE of $\hat{\mu}_{yw}$ in the presence of measurement errors is given by

$$\text{MSE}(\hat{\mu}_{yw}) = \theta(\sigma^2_y + \sigma^2_u) + \lambda(\sigma^2_{y_{(2)}} + \sigma^2_p) + G, \tag{19}$$

where $\theta = \frac{N-n}{Nn}$, $\lambda = \frac{N_2(f-1)}{Nn}$, and $G = \frac{W_2 f}{n}[\frac{[(\sigma^2_{y(2)}+\mu^2_{y(2)})\sigma^2_T+\sigma^2_s]W}{(\mu_T W+1-W)^2}]$.

2. A ratio estimator corresponding to the one in Gupta et al. [9] under modified HH is given by

$$\hat{\mu}_{rw} = \frac{\hat{\bar{y}}^*}{\bar{x}^*}\mu_x = \hat{R}^*_W\mu_x, \tag{20}$$

where $\hat{\bar{y}}^*$ is the ordinary mean estimator under modified HH and $\bar{x}^* = w_1\bar{x}_1 + w_2\bar{x}_2$ is the ordinary mean estimator under original HH procedure. The MSE of $\hat{\mu}_{rw}$ in the presence of measurement errors is given by

$$\begin{aligned}\text{MSE}^*(\hat{\mu}_{rw}) = {} & \theta(\sigma^2_y + R^2\sigma^2_x - 2R\rho_{yx}\sigma_y\sigma_x) + \lambda(\sigma^2_{y(2)} + R^2\sigma^2_{x(2)} - 2R\rho_{zx_{(2)}}\sigma_z\sigma_{x(2)}) \\ & + \theta(\sigma^2_u + R^2\sigma^2_v) + \lambda(\sigma^2_p + R^2\sigma^2_v) + G,\end{aligned} \tag{21}$$

where $R = \mu_y/\mu_x$ and $\rho_{zx_{(2)}} = \frac{\rho_{yx(2)}}{\sqrt{1+\frac{[\sigma^2_s+\sigma^2_T(\sigma^2_{y(2)}+\mu^2_{y(2)})]W}{\sigma^2_{y(2)}}}}$.

The MSE of $\hat{\mu}_{yw}$ and $\hat{\mu}_{rw}$, without measurement errors, may be obtained by putting $\sigma^2_v = \sigma^2_u = \sigma^2_p = 0$ in the above equations.

3.2 Proposed Mean Estimator

With this background, we use the generalized mean estimator considered in Khalil et al. [12, 13] but with non-response. This mean estimator includes a wide variety of mean estimators as special cases. The non-response version of this estimator is given by

$$\hat{\mu}_{pw} = (\hat{\bar{y}}^* + k(\mu_x - \bar{x}^*))\left(\frac{\bar{D}}{\bar{d}}\right)^v \tag{22}$$

where $\bar{d} = \phi(\alpha\bar{x}^* + \beta) + (1-\phi)(\alpha\mu_x + \beta)$, $\bar{D} = \alpha\mu_x + \beta$, k and v are suitable constants, and ϕ is assumed to be an unknown constant whose value is to be determined from optimality considerations. Also α and β are assumed to be some known parameters of the auxiliary variable X. Various estimators may be obtained by using different values of α and β. With $v = 1$, we get various regression-in-ratio estimators, and with $v = -1$, we get various regression-in-product estimators.

To obtain the MSE of this estimator, we define $\hat{\bar{y}}^* = \mu_y(1 + e_0^*)$ and $\bar{x}^* = \mu_x(1 + e_1^*)$ such that $E(e_0^*) = E(e_1^*) = 0$; $E(e_0^2) = \frac{1}{\mu_y^2}[\theta(\sigma_y^2 + \sigma_u^2) + \lambda(\sigma_{y_{(2)}}^2 + \sigma_p^2) + \frac{W_2 f}{n}[\frac{[(\sigma_{y_{(2)}}^2 + \mu_{y_{(2)}}^2)\sigma_T^2 + \sigma_s^2]W}{(\mu_T W + 1 - W)^2}]$; $E(e_1^{2*}) = \frac{1}{\mu_x^2}[\theta(\sigma_x^2 + \sigma_v^2) + \lambda(\sigma_{x_{(2)}^2} + \sigma_v^2)]$; $E(e_0^* e_1^*) = \theta\rho_{xy}\frac{\sigma_y}{\mu_y}\frac{\sigma_x}{\mu_x} + \lambda\rho_{zx(2)}\frac{\sigma_z}{\mu_z}\frac{\sigma_{x(2)}}{\mu_x}$, where $\rho_{zx_{(2)}} = \frac{\rho_{yx(2)}}{\sqrt{1 + \frac{[\sigma_s^2 + \sigma_T^2(\sigma_{y_{(2)}}^2 + \mu_{y_{(2)}}^2)]W}{\sigma_{y_{(2)}}^2}}}]$.

The bias of the proposed estimator, up to the second order of approximation, in the presence of measurement errors, is given by

$$\begin{aligned} Bias^*(\hat{\mu}_{pw}) \approx \theta\left[\left(kH + \frac{v+1}{v}\mu_y H^2\right)(\sigma_x^2 + \sigma_v^2) - H\rho_{yx}\sigma_y\sigma_x\right] \\ + \lambda\left[\left(kH + \frac{v+1}{v}\mu_y H^2\right)(\sigma_{x(2)}^2 + \sigma_v^2) - H\rho_{zx(2)}\sigma_z\sigma_{x(2)}\right], \end{aligned} \tag{23}$$

where $H = \frac{\alpha\phi v}{\alpha\mu_x + \beta}$. The bias of $\hat{\mu}_{pw}$, without measurement error, may be obtained by setting $\sigma_v^2 = 0$ in above equation.

Using Taylor's approximation up to the first order, we have

$$\hat{\mu}_{pw} - \mu_y \approx e_0^*\mu_y - k\mu_x e_1^* - H\mu_x\mu_y e_1^*. \tag{24}$$

Taking square and expectation in (24), we have

$$\begin{aligned} (\hat{\mu}_{pw} - \mu_y)^2 = e_0^{*2}\mu_y^2 + k^2\mu_x^2 e_1^{*2} + (H\mu_x\mu_y e_1^*)^2 - 2e_0^* e_1^* k\mu_x\mu_y - 2e_0^* e_1^* H\mu_x\mu_y^2 \\ + 2e_1^{*2} H\mu_x^2\mu_y, \end{aligned} \tag{25}$$

and

$$\begin{aligned} \mathrm{MSE}^*(\hat{\mu}_{pw}) &= E(\hat{\mu}_{pw} - \mu_y)^2 \\ &= \theta[\sigma_y^2 + (k + \phi v R_{pw})^2 \sigma_x^2 - 2(k + \phi v R_{pw}) \rho_{yx} \sigma_x \sigma_y] \\ &\quad + \lambda[\sigma_{y_(2)}^2 + (k + \phi v R_{pw})^2 \sigma_{x_(2)}^2 - 2(k + \phi v R_{pw}) \rho_{zx(2)} \sigma_x \sigma_z] \\ &\quad + \theta[\sigma_u^2 + (k + \phi v R_{pw})^2 \sigma_v^2] + \lambda[\sigma_p^2 + (k + \phi v R_{pw})^2 \sigma_v^2] + G \end{aligned} \tag{26}$$

where $R_{pw} = \frac{\alpha\mu_y}{\alpha\mu_x + \beta}$.

Minimization of the above expression (26) with respect to ϕ yields its optimum value as:

$$\phi_{opt} \cong \frac{\theta(\rho_{xy}\sigma_x\sigma_y - k(\sigma_x^2 + \sigma_v^2)) + \lambda(\rho_{zx_{(2)}}\sigma_z\sigma_{x_{(2)}} - k(\sigma_{x_{(2)}}^2 + \sigma_v^2))}{vR_{pw}[\theta(\sigma_x^2 + \sigma_v^2) + \lambda(\sigma_{x_{(2)}}^2 + \sigma_v^2)]}. \tag{27}$$

Substitution of ϕ_{opt} in $\mathrm{MSE}(\hat{\mu}_{pw})$ yields the minimum value as:

$$\begin{aligned} \mathrm{MSE}^*_{\min}(\hat{\mu}_{pw}) \cong{}& \theta(\sigma_y^2 + P^2\sigma_x^2 - 2P\rho_{yx}\sigma_x\sigma_y) + \lambda(\sigma_{y_{(2)}}^2 + P^2\sigma_{x_{(2)}}^2 - 2P\rho_{zx_{(2)}}\sigma_z\sigma_{x_{(2)}}) \\ &+ \theta(\sigma_u^2 + P^2\sigma_v^2) + \lambda(\sigma_p^2 + P^2\sigma_v^2) + G, \end{aligned} \tag{28}$$

where $P = \frac{\theta\rho_{yx}\sigma_x\sigma_y + \lambda\rho_{zx_{(2)}}\sigma_z\sigma_{x_{(2)}}}{\theta(\sigma_x^2 + \sigma_v^2) + \lambda(\sigma_{x_{(2)}}^2 + \sigma_v^2)}$.

The expression for the minimized MSE of the proposed estimator without ME may be obtained by putting $\sigma_u^2 = \sigma_v^2 = \sigma_p^2 = 0$ in the above expression, which gives

$$\mathrm{MSE}_{\min}(\hat{\mu}_{pw}) \cong \theta(\sigma_y^2 + P^2\sigma_x^2 - 2P\rho_{yx}\sigma_x\sigma_y) + \lambda(\sigma_{y_{(2)}}^2 + P^2\sigma_{x_{(2)}}^2 - 2P\rho_{zx_{(2)}}\sigma_z\sigma_{x_{(2)}}) + G. \tag{29}$$

Comparing the MSE expressions of $\hat{\mu}_{yw}$ in (19), $\hat{\mu}_{rw}$ in (21), and $\hat{\mu}_{pw}$ in (28) with measurement errors, it can be verified easily that

- $\mathrm{MSE}^*_{\min}(\hat{\mu}_{pw}) < \mathrm{MSE}^*(\hat{\mu}_{yw})$ if

$$-\frac{(\theta\rho_{yx}\sigma_x\sigma_y + \lambda\rho_{zx_{(2)}}\sigma_z\sigma_{x_{(2)}})^2}{\theta(\sigma_x^2 + \sigma_v^2) + \lambda(\sigma_{x_{(2)}}^2 + \sigma_v^2)} < 0, \tag{30}$$

- $\mathrm{MSE}^*_{\min}(\hat{\mu}_{pw}) < \mathrm{MSE}^*(\hat{\mu}_{rw})$ if

$$\frac{1}{2} - \frac{\mu_y}{2\mu_x}\frac{\theta(\sigma_x^2 + \sigma_v^2) + \lambda(\sigma_{x_{(2)}}^2 + \sigma_v^2)}{\theta\rho_{yx}\sigma_x\sigma_y + \lambda\rho_{zx_{(2)}}\sigma_z\sigma_{x_{(2)}}} < 1 \tag{31}$$

  and
- $\mathrm{MSE}^*(\hat{\mu}_{rw}) < \mathrm{MSE}^*(\hat{\mu}_{yw})$ if

$$\frac{\mu_y}{2\mu_x}\frac{\theta(\sigma_x^2+\sigma_v^2)+\lambda(\sigma_{x_{(2)}}^2+\sigma_v^2)}{\theta\rho_{yx}\sigma_x\sigma_y+\lambda\rho_{zx_{(2)}}\sigma_z\sigma_{x_{(2)}}}<1 \tag{32}$$

The conditions (30) and (31) always hold true. From (32), the ratio estimator is generally more efficient than the ordinary mean estimator if the measurement error on auxiliary variable X (σ_v^2) is small, and X and Y are strongly correlated.

4 Simulations

We will now compare the performance of the generalized mean estimator under simple random sampling with the other two estimators by a simulation study in this section. In the generalized mean estimator, we choose v and k to be 1, and ϕ to be its optimum value. In the simulation, ϕ is calculated by plugging the corresponding sample values in (27). As for α and β, we could use various parameters associated with the auxiliary variable such as the coefficient of variation (C_x) or kurtosis, but these choices do not impact the results in any meaningful way. As we can see in (28), minimized MSE is independent of α and β. Also, we ran extensive simulations and noticed that empirical MSEs also are almost the same for all choices of α and β. Therefore, we will only show the results where $\alpha = 1$ and $\beta = 0$. The scrambling variable S is taken to be a normal variate with mean equal to zero and variance, $\sigma_s^2 = 0.5 * \sigma_x^2$. T is also taken to be a normal variate but with mean equal to one and different variances. The measurement errors on X have a normal distribution with mean zero in both phases; the measurement errors of Y in the first phase and Z in the second phase have a normal distribution with mean zero. We use different variances (0, 5, 10) for measurement errors.

We consider a finite population of size 5000 generated from a bivariate normal distribution with means and covariance of (Y, X) as given below.

$$\textbf{Population}\quad \mu=\begin{bmatrix}10\\6\end{bmatrix},\quad \Sigma=\begin{bmatrix}16 & 9.051\\9.051 & 8\end{bmatrix},\quad \rho_{yx}=0.8$$

The parameters of the set of 5000 data points we generated using R are very close to the parameter values in (A) but not exactly same. For the simulation study, we use parameter values in (B) and not those in (A).

$$\mu_x = 6,\quad \sigma_x^2 = 8,\quad \mu_y = 10,\quad \sigma_y^2 = 16,\quad \rho_{yx} = 0.8 \tag{A}$$

$$\mu_x = 6.0228,\quad \sigma_x^2 = 8.1830,\quad \mu_y = 9.9864,\quad \sigma_y^2 = 16.1215,\quad \rho_{yx} = 0.8024 \tag{B}$$

We consider samples of size $n = 500$ using SRSWOR and assume a response rate of 40% in the first phase. This means in the first phase, only 200 (n_1) subjects provide a response to the survey question and 300 (n_2) of them do not. In the second phase, we take another sample ($n_s = \frac{n_2}{f}$) from the non-respondent group by using $f = 2, 3, 4$, respectively.

Coding for the simulations was done in R, and the results are averaged over 5000 iterations. The empirical MSE of the estimator $\hat{\mu}_y$ is computed by

$$\mathrm{MSE}^*(\hat{\mu}_w) = \frac{1}{5000}\sum_{i=1}^{5000}(\hat{\mu}_w - \mu)^2, \tag{33}$$

where $\hat{\mu}_w = \hat{\mu}_{yw}$, $\hat{\mu}_{rw}$, and $\hat{\mu}_{pw}$. Here, μ is the population mean of the sensitive study variable. The percent relative efficiency (PRE) of the estimator ($\hat{\mu}_w$) with respect to the ordinary mean estimator ($\hat{\mu}_{yw}$) is defined as

$$\mathrm{PRE} = \frac{\mathrm{MSE}^*(\hat{\mu}_{yw})}{\mathrm{MSE}^*(\hat{\mu}_w)} * 100. \tag{34}$$

We will also use the unified measure δ of the efficiency and the privacy as defined in Gupta et al. [10]. It is given by

$$\delta = \frac{\mathrm{MSE}^*(\hat{\mu}_w)}{\Delta_{DP}}. \tag{35}$$

In (35), MSE is used in place of Var(.) to account for biased estimators.

The simulation results are provided in the three tables below. In Table 1, we fix the response rate, Var(T), Var(S), and W but study the impact of vary the size of the measurement errors and the sampling fraction (f) in phase 2. In Table 2, we examine the impact of Var(T) and W.

These simulation results are discussed in Sect. 5.

5 Discussion

From the two tables, the empirical results are in good agreement with the corresponding theoretical results.

As the measurement errors increase, the MSE of each mean estimator increases. Also, the efficiency of each estimator gets worse as the value of f increases. For example in Table 1, the MSE of the generalized mean estimator increased from 0.1397 to 0.1804 as the variance of measurement errors increased from 1 to 10 when $f = 2$, and increased from 0.1601 to 0.3113 as the value of f increased from 2 to 4 when the variance of measurement error is 5. This is reasonable because larger measurement errors have larger negative impact on mean estimation and larger f value means we obtain smaller sample from the second call.

The results also showed that the MSEs of all mean estimators increase as W increases under non-response, both when measurement errors are present and when they are not present. For example, in Table 2, the MSE of the generalized mean estimator increased from 0.0966 to 0.1680 as the sensitivity level increased from 0.5 to 1 when variance of T is equal to 0.5. Therefore, optional RRT model

Table 1 Theoretical (bold) and empirical MSEs/PREs of the ORRT estimators with $\sigma_v^2 = \sigma_u^2 = \sigma_p^2 = 1, 5, 10$ when response rate = 40% , $W = 0.8$, $\sigma_T^2 = 0.5$ and $\sigma_s^2 = 0.5 * \sigma_x^2$.

Est.	f	Var(ME)	MSE			PRE		
			1	5	10	1	5	10
$\hat{\mu}_{yw}^{HH}$	2		**0.1680**	**0.1799**	**0.1948**	**100.0000**	**100.0000**	**100.0000**
			0.1679	0.1799	0.1951	100.0000	100.0000	100.0000
	3		**0.2470**	**0.2637**	**0.2847**	**100.0000**	**100.0000**	**100.0000**
			0.2397	0.2653	0.2854	100.0000	100.0000	100.0000
	4		**0.3261**	**0.3476**	**0.3745**	**100.0000**	**100.0000**	**100.0000**
			0.3173	0.3484	0.3756	100.0000	100.0000	100.0000
$\hat{\mu}_{rw}^{HH}$	2		**0.1514**	**0.1960**	**0.2518**	**110.9643**	**91.7857**	**77.3630**
			0.1498	0.1894	0.2407	112.0828	94.9842	81.0553
	3		**0.2236**	**0.2859**	**0.3638**	110.4651	**92.2350**	**78.2573**
			0.2223	0.2882	0.3611	107.8273	92.0541	79.0363
	4		**0.2957**	**0.3758**	**0.4758**	**110.2807**	**92.4960**	**78.7095**
			0.2854	0.3761	0.4705	111.1773	92.6349	79.8300
$\hat{\mu}_{pw}^{HH}$	2		**0.1397**	**0.1601**	**0.1804**	**120.2577**	**112.3673**	**107.9823**
			0.1392	0.1585	0.1790	120.6178	113.5016	108.9944
	3		**0.2071**	**0.2357**	**0.2642**	**119.2661**	**111.8795**	**107.7593**
			0.1972	0.2353	0.2626	121.5517	112.7497	108.6824
	4		**0.2745**	**0.3113**	**0.3480**	**118.7978**	**111.6608**	**107.6149**
			0.2653	0.3114	0.3486	119.6005	111.8818	107.7453

leads to better results than the non-optional model. Note that the model tends to become non-optional as W increases. Furthermore, the simple additive RRT model ($\sigma_T^2 = 0$) is more efficient in terms of PRE. But the general RRT model is better, if we examine the performance of various estimators with respect to the unified measure (δ) of efficiency and privacy. For instance, in Table 2, when the sensitivity level W is equal to 0.5, the MSE of the generalized mean estimator increases from 0.0284 to 0.1641 as the variance of T increases from 0 to 1, but the δ value decreases from 0.0069 to 0.0014.

It is clear from the theoretical conditions (30), (31), (32), and the simulation results that the generalized mean estimator is always more efficient than the ordinary RRT mean estimator and the ratio estimator, while the ratio estimator is less efficient than the ordinary mean estimator if the measurement errors on X are large. For example in Table 1, the MSE of the generalized mean estimator (0.1804) is less than the MSE of the ordinary mean estimator (0.1948) when the variance of measurement errors is 10 and the value of f is 2. However, the MSE of the ratio estimator (0.2518) is larger than the mean estimators because the measurement errors are large. This is because the ordinary mean estimator is not impacted by the measurement error in X. This was not so for the generalized estimator because the use of the regression term was able to overcome the measurement error burden due to X.

Table 2 Theoretical (bold) and empirical MSEs/PREs of the ORRT estimators when response rate = 40% , $\sigma_v^2 = \sigma_u^2 = \sigma_p^2 = 1$, $f = 2$ and $\sigma_s^2 = 0.5 * \sigma_x^2$

Estimator	W	Var(T)	MSE		PRE		δ
			Without ME	With ME	Without ME	With ME	
$\hat{\mu}_{yw}^{HH}$	0.5	0	**0.0537**	**0.0567**	**100.0000**	**100.0000**	**0.0139**
			0.0545	0.0584	100.0000	100.0000	0.0143
		0.5	**0.1219**	**0.1249**	**100.0000**	**100.0000**	**0.0020**
			0.1172	0.1196	100.0000	100.0000	0.0020
		1	**0.1895**	**0.1924**	**100.0000**	**100.0000**	**0.0016**
			0.1820	0.1843	100.0000	100.0000	0.0016
	0.8	0	**0.0566**	**0.0596**	**100.0000**	**100.0000**	**0.0146**
			0.0559	0.0598	100.0000	100.0000	0.0146
		0.5	**0.1650**	**0.1680**	**100.0000**	**100.0000**	**0.0027**
			0.1651	0.1679	100.0000	100.0000	0.0027
		1	**0.2716**	**0.2746**	**100.0000**	**100.0000**	**0.0023**
			0.2746	0.2771	100.0000	100.0000	0.0023
	1	0	**0.0586**	**0.0616**	**100.0000**	**100.0000**	**0.0151**
			0.0587	0.0625	100.0000	100.0000	0.0153
		0.5	**0.1933**	**0.1963**	**100.0000**	**100.0000**	**0.0032**
			0.1847	0.1916	100.0000	100.0000	0.0031
		1	**0.3255**	**0.3285**	**100.0000**	**100.0000**	**0.0028**
			0.3225	0.3226	100.0000	100.0000	0.0027
$\hat{\mu}_{rw}^{HH}$	0.5	0	**0.0290**	**0.0403**	**185.1724**	**140.6948**	**0.0099**
			0.0301	0.0420	181.0631	139.0476	0.0103
		0.5	**0.0972**	**0.1084**	**125.4115**	**115.2214**	**0.0018**
			0.0940	0.1028	124.6809	116.3424	0.0017
		1	**0.1648**	**0.1759**	**114.9879**	**109.3803**	**0.0015**
			0.1597	0.1679	113.9637	109.7677	0.0014
	0.8	0	**0.0319**	**0.0432**	**177.4295**	**137.9630**	**0.0106**
			0.0330	0.0454	169.3939	131.7181	0.0111
		0.5	**0.1403**	**0.1514**	**117.6051**	**110.9643**	**0.0025**
			0.1425	0.1498	115.8596	112.0828	0.0024
		1	**0.2469**	**0.2581**	**110.0041**	**106.3929**	**0.0022**
			0.2531	0.2588	108.4947	107.0711	0.0022
	1	0	**0.0339**	**0.0451**	**172.8614**	**136.5854**	**0.0110**
			0.0345	0.0451	170.1449	138.5809	0.0110
		0.5	**0.1686**	**0.1798**	**114.6501**	**109.1769**	**0.0029**
			0.1640	0.1741	112.6220	110.0517	0.0028
		1	**0.3008**	**0.3119**	**108.2114**	**105.3222**	**0.0026**
			0.2940	0.3058	109.6939	105.4938	0.0026

Table 2 (continued)

Estimator	W	Var(T)	MSE		PRE		δ
			Without ME	With ME	Without ME	With ME	
$\hat{\mu}_{pw}^{HH}$	0.5	0	**0.0220**	**0.0284**	**244.0909**	**199.6479**	**0.0069**
			0.0228	0.0296	239.0351	197.2973	0.0072
		0.5	**0.0903**	**0.0966**	**134.9945**	**129.2961**	**0.0016**
			0.0864	0.0912	135.6481	131.1404	0.0015
		1	**0.1578**	**0.1641**	**120.0887**	**117.2456**	**0.0014**
			0.1518	0.1562	119.8946	117.9898	0.0013
	0.8	0	**0.0250**	**0.0313**	**226.4000**	**190.4153**	**0.0077**
			0.0252	0.0322	221.8254	185.7143	0.0079
		0.5	**0.1333**	**0.1397**	**123.7809**	**120.2577**	**0.0023**
			0.1348	0.1392	122.4777	120.6178	0.0023
		1	**0.2399**	**0.2463**	**113.2138**	**111.4901**	**0.0021**
			0.2451	0.2485	112.0359	111.5091	0.0021
	1	0	**0.0269**	**0.0333**	**217.8439**	**184.9850**	**0.0081**
			0.0276	0.0339	212.6812	184.3658	0.0083
		0.5	**0.1617**	**0.1680**	**119.5424**	**116.8452**	**0.0027**
			0.1553	0.1629	118.9311	117.6182	0.0027
		1	**0.2938**	**0.3002**	**110.7897**	**109.4270**	**0.0025**
			0.2847	0.2944	113.2771	109.5788	0.0025

Acknowledgements The authors express very sincere gratitude to the reviewers for their constructive suggestions which helped improve the presentation of the paper.

References

1. Ahmed S, Shabbir J, Gupta S (2017) Use of scrambled response model in estimating the finite population mean in presence of non-response when coefficient of variation is known. Commun Stat Theory Methods 46(17):8435–8449
2. Audu A, Singh R, Khare S, Dauran NS (2020) Almost unbiased estimators for population mean in the presence of non-response and measurement error. J Stat Manag Syst. https://doi.org/10.1080/09720510.2020.1759209
3. Bhushan S, Pandey AP (2019) An efficient estimation of population mean under non-response. Commun Stat Appl Methods 26:11–25
4. Bhushan S, Pandey AP (2020) An efficient estimation procedure for the population mean under non-response. Statistica 79(4):363–378
5. Diana G, Perri PF (2011) A class of estimators for quantitative sensitive data. Stat Pap 52:633–650
6. Diana G, Riaz S, Shabbir J (2014) Hansen and Hurwitz estimator with scrambled response on the second call. J Appl Stat 41(3):596–611

7. Gupta S, Gupta B, Singh S (2002) Estimation of sensitivity level of personal interview survey questions. J Stat Plan Inference 100:239–247
8. Gupta S, Shabbir J, Sousa R, Corte-Real P (2012) Estimation of the mean of a sensitive variable in the presence of auxiliary information. Commun Stat Theory Methods 41(13–14):2394–2404
9. Gupta S, Kalucha J, Shabbir J, Dass BK (2014) Estimation of finite population mean using optional RRT models in the presence of nonsensitive auxiliary information. Am J Math Manag Sci 33(2):147–159
10. Gupta S, Mehta S, Shabbir J, Khalil S (2018) A unified measure of respondent privacy and model efficiency in quantitative RRT models. J Stat Theory Pract 12(3):506–511
11. Hansen MH, Hurwitz WN (1946) The problem of non-response in sample surveys. J Am Stat Assoc 41:517–529
12. Khalil S, Noor-ul-Amin M, Hanif M (2018) Estimation of population mean for a sensitive variable in the presence of measurement error. J Stat Manag Syst 21(1):81–91
13. Khalil S, Zhang Q, Gupta S (2018) Mean estimation of sensitive variables under measurement errors using optional RRT models. Commun Stat Simul Comput. https://doi.org/10.1080/03610918.2019.1584298
14. Khare BB, Srivastava S (1997) Transformed ratio type estimators for the population mean in the presence of non-response. Commun Stat Theory Methods 26(5):1779–1791
15. Khare BB, Sinha RR (2007) Estimation of ratio of the two population means using multi-auxiliary characters in the presence of non-response. Narosa Publishing House 10(3):3–14
16. Khare BB, Sinha RR (2009) On class of estimators for population mean using multi-auxiliary characters in the presence of non-response. Appl Math Comput 10(3):45–56
17. Khare BB, Sinha RR (2011) Estimation of population mean using multi-auxiliary characters with subsampling the nonrespondents. Stat Transit 12(1):3–14
18. Kumar M, Singh R, Singh AK, Smarandache F (2011) Some ratio type estimators under measurement errors. World Appl Sci J 14(2):272–276
19. Kumar S, Singh HP, Bhougal S, Gupta R (2011) A class of ratio-cum-product type estimators under double sampling in the presence of non- response. J Math Stat 40(4):589–599
20. Kumar M, Singh R, Sawan N, Chauhan P (2011) Exponential ratio method of estimators in the presence of measurement errors. Int J Agric Stat Sci 7(2):457–461
21. Makhdum M, Sanaullah A, Hanif M (2020) A modified regression-cum-ratio estimator of population mean of a sensitive variable in the presence of non-response in simple random sampling. J Stat Manag Syst 23(3):495–510
22. Rao PSRS (1986) Ratio estimation with sub-sampling the non-respondents. Survey Method 12(2):217–230
23. Singh SR, Sharma P (2015) Method of estimation in the presence of non-response and measurement errors simultaneously. J Mod Appl Stat Methods 14(1):107–121
24. Singh N, Vishwakarma GK, Kim JM (2019) Computing the effect of measurement error on efficient variant of ratio and product estimator using auxiliary variable. Commun Stat Simul Comput. https://doi.org/10.1080/03610918.2019.1656742
25. Singh N, Vishwakarma GK (2019) A generalized class of estimator of population mean with the combined effect of measurement errors and non-response in sample survey. Rev Investig Oper 40:275–285
26. Unal C, Kadilar C (2020) Exponential type estimator for the population mean in the presence of non-response. J Stat Manag Syst 23(3):603–615
27. Yan Z, Wang J, Lai J (2008) An efficiency and protection degree-based comparison among the quantitative randomized response strategies. Commun Stat Theory Methods 38(3):400–408
28. Yaqub M, Shabbir J, Gupta SN (2017) Estimation of population mean based on dual use of auxiliary information in non response. Commun Stat Theory Methods 46(24):12130–12151