RESEARCH ARTICLE

Estimation of finite population mean for a sensitive variable using dual auxiliary information in the presence of measurement errors

Erum Zahid^{*}, Javid Shabbir

Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan

* erumzahid22@gmail.com

Abstract

In this study, we propose a new improved estimator of population mean for the sensitive variable in the presence of measurement error under simple and stratified random sampling. This estimator accounts the auxiliary information as well as the ranks of the auxiliary variable. From theoretical and numerical studies it is shown that a new improved estimator performs better than the existing estimators under study.

OPEN ACCESS

Citation: Zahid E, Shabbir J (2019) Estimation of finite population mean for a sensitive variable using dual auxiliary information in the presence of measurement errors. PLoS ONE 14(2): e0212111. <https://doi.org/10.1371/journal.pone.0212111>

Editor: Lei Shi, Yunnan University of Finance and Economics, CHINA

Received: May 29, 2018

Accepted: January 28, 2019

Published: February 11, 2019

Copyright: © 2019 Zahid, Shabbir. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This study was supported by a fellowship from the Higher Education Commission (HEC 213-63287-2SS2-187) to EZ. The funder had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing interests: "The authors have declared that no competing interests exist."

1 Introduction

In survey sampling, the assumption is made that all the observations are carefully considered on the characteristics under study so the information we obtained is error free. But in practice this assumption is not achieved due to many reasons, including non-response which may arises due to refusal of respondents to give the information or not at home or lack of interest or due to some sensitive issues. In analysis, a basic assumption is that all observations are measured correctly. In multiple regression model, it is assumed that all observations based on the study variable and the auxiliary variable are observed without any error. In many situations these assumptions are violated because of the following reasons. (i) Under the context of qualitative, it is hard to measure some variables (e.g., intelligence, taste, ability, climate, education, poverty etc.). So we use the dummy variables and observations are recorded in terms of values of dummy variables. (ii) In application, some variables are clearly defined but it is hard to take the correct observations (e.g., age is either under reported or over reported in complete year). (iii) It is no doubt that some variables are conceptually defined but is hard to take correct observation on it, instead the observations are taken on closely related variables (e.g., level of education is measured by the number of years of schooling). In all above mentioned cases, it is not possible to obtain true value of the variable. Instead it is recorded with error. So measurement error (ME) appeared because of difference between observed and true value. Also ME is due to the use of imperfect measure of true values of variables. Suppose we are interested to get the average level of anxiety among students, So we take a random sample of some students and measure their level of anxiety. Then we calculate the mean level of anxiety i.e. sample mean.

The normality assumption says that if you repeat this process many times and plot the sample means, the distribution will be normal. Usually measurement error and randomized response are studied separately using the known auxiliary or additional information. In reality, when the variable of interest is sensitive, the respondents hesitate to provide the personal information, which gives rise to measurement error.

To estimate the population mean, few researchers discussed the problem of measurement error. [1] discussed some important sources of measurement error in survey data. [2] done the estimation of population mean in the presence of measurement error for ratio-product type estimators. [3] and [4] presented the ratio method of estimation in the presence of measurement error. Further the work is extended by [5]. [6], [7] and [8] studied measurement error and non response together. [9] suggested an estimator for the estimation of population mean in the presence of measurement error and non response under stratified random sampling.

In survey sampling, when the variable of interest is sensitive, then the respondents hesitate to provide their personal information. Direct survey on sensitive question increases the relative bias. [10] introduced the randomized response technique (RRT), which reduces the possible bias and is used to obtain the true information while insuring the privacy of the respondents. For estimation of mean of a sensitive quantitative variable the Randomized Response model (RRM) is extended by [11]. [12] introduced the scrambled randomized response method. [13] proposed the optional RRT method and further this work is extended by [14]. [15] used the scrambled response technique for the estimation of population mean when coefficient of variation is known. [16] used the empirical Bayes estimation for the estimation of sensitive variable. [17] studied the estimation of population mean of sensitive variable in the presence of nonsensitive auxiliary information. [18] and [19] studied the improved estimation of population mean in simple and stratified random sampling.

When the correlation between the study variable and the auxiliary variable is sufficient, then the ranks of the auxiliary variable are also correlated with the study variable and consequently the precision of the estimator increased. [20] suggested the concept of ranks of the auxiliary variable to make efficient estimates. In practice, not much literature has been found in estimating the population mean for the sensitive variable in the presence of measurement error based on dual use of the auxiliary information.

The present paper is organized as: Section 2 gives existing estimators and an improved proposed estimator of population mean for sensitive variable in the presence of measurement error under simple random sampling. Both theoretical and numerical comparison are done in Section 2. In Section 3, some existing estimators and an improved class of estimators is suggested for estimating the finite population mean by incorporating both measurement error and sensitive information simultaneously under stratified random sampling. Efficiency comparison, numerical results and simulation study are also presented in Section 3. Conclusion is given in Section 4.

2 Estimators under simple random sampling

Let $\Omega = \Omega_1, \Omega_2, \dots, \Omega_N$ be a finite population of size N . Suppose that a simple random sample of size n is drawn from Ω by using simple random sampling without replacement. Let Y be the sensitive study variable, which is not observed directly and X be the non-sensitive auxiliary variable which has positive correlation with Y . Let R_x be ranks of the auxiliary variable X . Let S be a scrambling variable which is independent of Y and X . We assume that S has zero mean and variance S_y^2 . The respondent is asked to give a scrambled response for the study variable Y given by $Z = Y + S$ and in addition asked to provide a true response for X .

Let $(x_i, r_{x,i}, y_i, z_i)$ be the observed values and $(X_i, R_{x,i}, Y_i, Z_i)$ be the actual values on the variables (X, R_x, Y, Z) respectively. Then the measurement errors be $V_i = x_i - X_i$, $U_i = z_i - Z_i$ and $T_i = r_{x,i} - R_{x,i}$. These measurement errors are assumed to be uncorrelated having normal distribution with zero mean and variances S_V^2 , S_U^2 and S_T^2 respectively. Let S_X^2 , $S_{R_x}^2$ and S_Z^2 be the population variances; ρ_{XZ} , ρ_{XR_x} and ρ_{ZR_x} be the coefficients of correlation between their subscripts.

2.1 Existing estimators in literature

In this section we consider the following existing estimators.

Mean estimator. The usual unbiased mean per unit estimator, is given by

$$\bar{y}_0 = \bar{z}, \quad (1)$$

where $\bar{z}$ is given in Eq (12). The variance of $\bar{y}_0$, is given by

$$Var(\bar{y}_0) = \lambda(S_Z^2 + S_U^2), \quad (2)$$

where $\lambda = (n^{-1} - N^{-1})$.

Ratio estimator. The traditional ratio estimator, is given by,

$$\bar{y}_R = \bar{z} \left(\frac{\bar{X}}{\bar{x}} \right), \quad (3)$$

where $\bar{x}$ is the sample mean (see Eq (13)) and $\bar{X}$ is known population mean. The bias and mean square error of $\bar{y}_R$ to first degree of approximation, are given by

$$B(\bar{y}_R) \cong \frac{\lambda}{\bar{X}} [R(S_X^2 + S_V^2) - \rho_{XZ} S_X S_Z] \quad (4)$$

and

$$MSE(\bar{y}_R) \cong \lambda[(S_Z^2 + S_U^2) + R^2(S_X^2 + S_V^2) - 2R\rho_{XZ} S_X S_Z]. \quad (5)$$

where $R = \frac{\bar{Z}}{\bar{X}}$.

Difference estimator. The usual difference estimator is given by,

$$\bar{y}_D = \bar{z} + d(\bar{X} - \bar{x}), \quad (6)$$

where d is the constant, whose value is to be determined optimally. The minimum variance of $\bar{y}_D$, is given by

$$Var(\bar{y}_D)_{min} = \lambda \left[S_Z^2 + S_U^2 - \frac{\rho_{XZ}^2 S_Z^2 S_X^2}{(S_X^2 + S_V^2)} \right], \quad (7)$$

where optimum value of d is $d_{(opt)} = \frac{\rho_{XZ} S_Z S_X}{(S_X^2 + S_V^2)}$.

Khalil estimator. Recently [21] proposed the generalized randomized response estimator, given by,

$$\bar{y}_K = [\bar{z} + k(\bar{X} - \bar{x})] \left(\frac{\bar{W}}{\bar{w}} \right)^g, \quad (8)$$

where $\bar{w} = \phi(\alpha\bar{x} + \gamma) + (1 - \phi)(\alpha\bar{X} + \gamma)$ and $\bar{W} = \alpha\bar{X} + \gamma$;

k and g are constants, and ϕ is assumed to be an unknown constant which is determined optimally as $\phi = \frac{\rho_{XZ} S_Z S_X - k(S_X^2 + S_V^2)}{gR(S_X^2 + S_V^2)}$. Also $\alpha(\neq 0)$ and γ are assumed to be some known parameters

of the auxiliary variable X . The bias and minimum MSE of $\bar{y}_K$ to first degree approximation, are given by

$$B(\bar{y}_K) \cong \frac{\lambda}{Z} \left[\frac{g(g+1)}{2} \phi^2 R^2 (S_x^2 + S_y^2) - g\phi R \rho_{xz} S_x S_z + g\phi k R (S_x^2 + S_y^2) \right] \quad (9)$$

and

$$MSE(\bar{y}_K)_{min} \cong \lambda \left[S_z^2 + S_U^2 - \frac{\rho_{xz}^2 S_z^2 S_x^2}{(S_x^2 + S_y^2)} \right], \quad (10)$$

which is exactly equal to the variance of the difference estimator $\bar{y}_D$, but $\bar{y}_D$ is preferable over $\bar{y}_K$ because of unbiasedness.

2.2 The proposed estimator

We propose an improved randomized response estimator for estimating the population mean of the sensitive variable, dealing with the problem of measurement error. Measurement error is considered on both the study and the auxiliary variables. A scrambled response of Y is observed in form of $Z = Y + S$, where S is distributed as $N(0, S_s^2)$. The proposed estimator, is given by

$$\bar{y}_P = \bar{z} + m_1 (\bar{X} - \bar{x}) \frac{\bar{X}}{\bar{x}} + m_2 (\bar{R}_x - \bar{r}_x) \frac{\bar{R}_x}{\bar{r}_x}, \quad (11)$$

where, m_1 and m_2 are constants whose values are to be determined. For obtaining the bias and mean square error, we assume that

$$\delta_Z = \sum_{i=1}^n (Z_i - \bar{Z}), \delta_U = \sum_{i=1}^n U_i.$$

$$\delta_X = \sum_{i=1}^n (X_i - \bar{X}), \delta_V = \sum_{i=1}^n V_i.$$

$$\delta_{R_x} = \sum_{i=1}^n (R_{x_i} - \bar{R}_x), \delta_T = \sum_{i=1}^n T_i.$$

Adding δ_Z and δ_U , we get

$$\delta_Z + \delta_U = \sum_{i=1}^n (Z_i - \bar{Z}) + \sum_{i=1}^n U_i.$$

Dividing both sides by n , and then simplifying, we get

$$\bar{z} = \bar{Z} + \frac{1}{n} (\delta_Z + \delta_U). \quad (12)$$

Similarly, we can write

$$\bar{x} = \bar{X} + \frac{1}{n} (\delta_X + \delta_V), \quad (13)$$

and

$$\bar{r}_x = \bar{R}_x + \frac{1}{n}(\delta_{R_x} + \delta_T). \quad (14)$$

Let $W_Z = \frac{\delta_Z + \delta_U}{n}$, $W_X = \frac{\delta_X + \delta_V}{n}$ and $W_{R_x} = \frac{\delta_{R_x} + \delta_T}{n}$. In order to get the bias and MSE of the proposed estimator, we consider the following relative error terms:

$$\text{Let } e_0 = \frac{\bar{z} - \bar{Z}}{\bar{Z}} = \frac{W_Z}{\bar{Z}}, e_1 = \frac{\bar{x} - \bar{X}}{\bar{X}} = \frac{W_X}{\bar{X}}, e_2 = \frac{\bar{r}_x - \bar{R}_x}{\bar{R}_x} = \frac{W_{R_x}}{\bar{R}_x}, E(e_j) = 0, j = 0, 1, 2. E(e_0^2) = \frac{\lambda(S_Z^2 + S_U^2)}{\bar{Z}^2},$$

$$E(e_1^2) = \frac{\lambda(S_X^2 + S_V^2)}{\bar{X}^2}, E(e_2^2) = \frac{\lambda(S_{R_x}^2 + S_T^2)}{\bar{R}_x^2}, E(e_0 e_1) = \frac{\lambda \rho_{XZ} S_X S_Z}{\bar{X} \bar{Z}}, E(e_0 e_2) = \frac{\lambda \rho_{ZR_x} S_Z S_{R_x}}{\bar{Z} \bar{R}_x} \text{ and } E(e_1 e_2) = \frac{\lambda \rho_{XR_x} S_X S_{R_x}}{\bar{X} \bar{R}_x}$$

Solving Eq (11) in terms of errors, we have

$$\bar{y}_{(p)} = \bar{Z}(1 + e_0) - m_1 \bar{X} \left(\frac{e_1}{1 + e_1} \right) - m_2 \bar{R}_x \left(\frac{e_2}{1 + e_2} \right), \quad (15)$$

Further simplifying, and keeping the terms up to power 2, we have

$$\bar{y}_{(p)} - \bar{Z} \cong \bar{Z}e_0 - m_1 \bar{X}e_1 + m_1 \bar{X}e_1^2 - m_2 \bar{R}_x e_2 + m_2 \bar{R}_x e_2^2, \quad (16)$$

On the lines of [22] and [23], we use the approximation method to derive the MSE of our proposed estimator in simple and stratified random sampling. The signal to noise ratio can easily be obtained by using the expression $C.V = \frac{S.D}{Mean}$. Using above equation the bias of $\bar{y}_{(p)}$, is given by

$$B(\bar{y}_{(p)}) \cong m_1 \frac{\lambda(S_X^2 + S_V^2)}{\bar{X}} + m_2 \frac{\lambda(S_{R_x}^2 + S_T^2)}{\bar{R}_x}. \quad (17)$$

Squaring and taking expectation in Eq (16), we have

$$MSE(\bar{y}_{(p)}) \cong \lambda(S_Z^2 + S_U^2) + m_1^2 \lambda(S_X^2 + S_V^2) + m_2^2 \lambda(S_{R_x}^2 + S_T^2) \\ + 2m_1 m_2 \lambda \rho_{XR_x} S_X S_{R_x} - 2m_1 \lambda \rho_{XZ} S_X S_Z - 2m_2 \lambda \rho_{ZR_x} S_Z S_{R_x}. \quad (18)$$

The optimum values of m_1 and m_2 are

$$m_{1(opt)} = \frac{\lambda^2(\rho_{XR_x} S_X S_{R_x})(\rho_{ZR_x} S_Z S_{R_x}) - \lambda^2(S_{R_x}^2 + S_T^2)(\rho_{XZ} S_X S_Z)}{\lambda^2(\rho_{XR_x} S_X S_{R_x})^2 - \lambda^2(S_{R_x}^2 + S_T^2)(S_X^2 + S_V^2)}, \quad (19)$$

and

$$m_{2(opt)} = \frac{\lambda^2(\rho_{XR_x} S_X S_{R_x})(\rho_{XZ} S_X S_Z) - \lambda^2(S_X^2 + S_V^2)(\rho_{ZR_x} S_Z S_{R_x})}{\lambda^2(\rho_{XR_x} S_X S_{R_x})^2 - \lambda^2(S_{R_x}^2 + S_T^2)(S_X^2 + S_V^2)}, \quad (20)$$

Substitute the optimum values of m_1 and m_2 in Eq (18), we get the minimum MSE of $\bar{y}_p$, given by

$$MSE(\bar{y}_p)_{min} \cong \frac{Q}{\lambda^2(\rho_{XR_x} S_X S_{R_x})^2 - \lambda^2(S_{R_x}^2 + S_T^2)(S_X^2 + S_V^2)}, \quad (21)$$

where,

$$Q = \lambda^3(S_Z^2 + S_U^2)(\rho_{XR_x} S_X S_{R_x})^2 + \lambda^3(S_{R_x}^2 + S_T^2)(\rho_{XZ} S_X S_Z)^2 \\ + \lambda^3(S_X^2 + S_V^2)(\rho_{ZR_x} S_Z S_{R_x})^2 - \lambda^3(S_Z^2 + S_U^2)(S_X^2 + S_V^2)(S_{R_x}^2 + S_T^2) \\ - 2\lambda^3(\rho_{XZ} S_X S_Z)(\rho_{XR_x} S_X S_{R_x})(\rho_{ZR_x} S_Z S_{R_x}).$$

2.3 Efficiency comparison

We compare the proposed estimator $\bar{y}_p$ with respect to $\bar{y}_0$, $\bar{y}_R$, $\bar{y}_D$ and $\bar{y}_K$, given by

1. From Eqs (2) and (21)

$$MSE(\bar{y}_p)_{min} < Var(\bar{y}_0), \text{ if } \frac{Q}{\lambda^2(\rho_{XR_x} S_X S_{R_x})^2 - \lambda^2(S_{R_x}^2 + S_T^2)(S_X^2 + S_V^2)} - \lambda(S_Z^2 + S_U^2) < 0$$

2. From Eqs (5) and (21)

$$MSE(\bar{y}_p)_{min} < MSE(\bar{y}_R), \text{ if } \frac{Q}{\lambda^2(\rho_{XR_x} S_X S_{R_x})^2 - \lambda^2(S_{R_x}^2 + S_T^2)(S_X^2 + S_V^2)} - \lambda((S_Z^2 + S_U^2) + R^2(S_X^2 + S_V^2) - 2R\rho_{XZ} S_X S_Z) < 0$$

3. From Eqs (7) and (21)

$$MSE(\bar{y}_p)_{min} < Var(\bar{y}_D)_{min}, \text{ if } \frac{Q}{\lambda^2(\rho_{XR_x} S_X S_{R_x})^2 - \lambda^2(S_{R_x}^2 + S_T^2)(S_X^2 + S_V^2)} - \lambda\left(S_Z^2 + S_U^2 - \frac{\rho_{XZ}^2 S_Z^2 S_X^2}{(S_X^2 + S_V^2)}\right) < 0$$

4. From Eqs (10) and (21)

$$MSE(\bar{y}_p)_{min} < MSE(\bar{y}_K)_{min}, \text{ if } \frac{Q}{\lambda^2(\rho_{XR_x} S_X S_{R_x})^2 - \lambda^2(S_{R_x}^2 + S_T^2)(S_X^2 + S_V^2)} - \lambda\left(S_Z^2 + S_U^2 - \frac{\rho_{XZ}^2 S_Z^2 S_X^2}{(S_X^2 + S_V^2)}\right) < 0$$

The proposed class of estimator $\bar{y}_p$ is more efficient than other existing estimators when above conditions 1 to 4 are satisfied.

2.4 Numerical results

In this section two populations are generated for simulation study and two are based on real data sets.

2.4.1 Simulation study. We have generated two populations of size 1,000 from multivariate normal distribution with different covariance matrices. The results of simulation is given in Tables 1 and 2. The population means and covariance matrices, are given below:

Population I

$$\mu_1 = \begin{bmatrix} 500 \\ 500 \\ 500 \end{bmatrix}$$

Table 1. MSE of different estimators for Population I under simulation.

Estimators	MSE without measurement error			MSE with measurement error		
	<i>n</i>			<i>n</i>		
	100	150	200	100	150	200
$\bar{y}_0$	9.469961	5.962568	4.208872	9.471728	5.963726	4.209671
$\bar{y}_R$	2.102341	1.323696	0.934374	2.106382	1.326300	0.935956
$\bar{y}_D, \bar{y}_K$	2.088259	1.314829	0.928115	2.092113	1.317314	0.929633
$\bar{y}_p$	1.924486	1.211713	0.855327	1.927487	1.213608	0.856596

<https://doi.org/10.1371/journal.pone.0212111.t001>

Table 2. MSE of different estimators for Population II under simulation.

Estimators	MSE without measurement error			MSE with measurement error		
	<i>n</i>			<i>n</i>		
	100	150	200	100	150	200
$\bar{y}_0$	3.732838	2.350305	1.659039	3.734604	2.351463	1.659838
$\bar{y}_R$	3.511012	2.210637	1.560449	3.515051	2.213239	1.562031
$\bar{y}_D, \bar{y}_K$	2.445976	1.540059	1.087100	2.448365	1.541613	1.088114
$\bar{y}_P$	2.250465	1.416959	1.000206	2.252707	1.418413	1.001175

<https://doi.org/10.1371/journal.pone.0212111.t002>

and

$$\sum_1 = \begin{bmatrix} 1000 & 800 & 810 \\ 800 & 850 & 820 \\ 810 & 820 & 840 \end{bmatrix}$$

$$\rho_{XY} = 0.8820, \rho_{XR_x} = 0.9702 \text{ and } \rho_{YR_x} = 0.8884$$

Population II

$$\mu_2 = \begin{bmatrix} 500 \\ 500 \\ 500 \end{bmatrix}$$

and

$$\sum_2 = \begin{bmatrix} 400 & 270 & 220 \\ 270 & 500 & 300 \\ 220 & 300 & 675 \end{bmatrix}$$

$$\rho_{XY} = 0.5897, \rho_{XR_x} = 0.5105 \text{ and } \rho_{YR_x} = 0.4965$$

Covariance matrices shows the distribution of sensitive variable Y , the auxiliary variable X and the ranks of the auxiliary variable R_x . There is high correlation in Population I, and weak correlation in Population II. The scrambling response S is distributed as $N(0, 0.01\sigma_X)$. The response variable is $Z = Y + S$. We estimate the MSE using $k = 1000$ samples of various sizes selected from each population. Three different sample sizes $n = 100, 150, 200$ are taken from both populations. The expression is given below:

$$MSE(\bar{y}_i) = \frac{1}{k} \sum_{j=1}^k (y_{ij} - \bar{y})^2,$$

where $i = 0, R, D, K, P$.

Tables 1 and 2 show that the proposed estimator $\bar{y}_P$ performs better as compared to all other existing estimators for both populations. The MSE of proposed estimators is smaller for Population I as compared to Population II because there is high correlation between the variables in Population I as compared to Population II. As the sample size increases MSE of all the

Table 3. Data summary.

Variable	Mean	St.Dev	Min	Med	Max	C.V
Forced expiratory volume (Y)	2.63	0.86	0.79	2.54	5.79	0.32
Age (X)	9.93	2.95	3.00	10.00	19.00	0.29
Smoke (S) 0,1	0.10	0.30	0.00	0.00	1.00	3.00

<https://doi.org/10.1371/journal.pone.0212111.t003>

Table 4. Data summary.

Variable	Mean	St.Dev	Min	Med	Max	C.V
Forced expiratory volume (Y)	2.63	0.86	0.79	2.54	5.79	0.32
Age (X)	9.93	2.95	3.00	10.00	19.00	0.29
Sex (S)0,1	0.51	0.50	0.00	1.00	1.00	0.98

<https://doi.org/10.1371/journal.pone.0212111.t004>

estimators decreases, and it is observed that MSEs of both difference estimator $\bar{y}_D$ and Khalil estimator $\bar{y}_K$ is same, but $\bar{y}_D$ is preferable over $\bar{y}_K$ because of unbiasedness.

2.4.2 Application to real data. In this section we have considered two data sets for numerical comparisons. Both data sets consist of 654 observations. The data summary is given below (see Tables 3 and 4) and results are given in Tables 5 and 6.

Population III (Source: [24])

Population IV (Source: [24])

In both populations the study and the auxiliary variables are identical, but scrambling responses are different. The correlation coefficients for both the Populations are: $\rho_{XY} = 0.7564$, $\rho_{XR_x} = 0.7831$ and $\rho_{YR_x} = 0.6161$. In Population, III and IV smoke (No = 0, Yes = 1) and sex (Female = 0, Male = 1) are taken as scrambling responses respectively.

Table 5. MSE of different estimators for Population III under real data.

Estimators	MSE without measurement error			MSE with measurement error		
	<i>n</i>			<i>n</i>		
	65	100	130	65	100	130
$\bar{y}_0$	0.011658	0.007127	0.005186	0.021543	0.014154	0.010172
$\bar{y}_R$	0.004445	0.002717	0.001977	0.014575	0.009893	0.007074
$\bar{y}_D, \bar{y}_K$	0.004395	0.002687	0.001955	0.014520	0.009859	0.007049
$\bar{y}_P$	0.004364	0.002668	0.001941	0.014412	0.009794	0.007001

<https://doi.org/10.1371/journal.pone.0212111.t005>

Table 6. MSE of different estimators for Population IV under real data.

Estimators	MSE without measurement error			MSE with measurement error		
	<i>n</i>			<i>n</i>		
	65	100	130	65	100	130
$\bar{y}_0$	0.013883	0.008487	0.006175	0.017534	0.010651	0.007757
$\bar{y}_R$	0.008993	0.005498	0.004000	0.012875	0.007820	0.005696
$\bar{y}_D, \bar{y}_K$	0.008606	0.005261	0.003828	0.012426	0.007520	0.005493
$\bar{y}_P$	0.008600	0.005258	0.003825	0.012374	0.007505	0.005467

<https://doi.org/10.1371/journal.pone.0212111.t006>

Tables 5 and 6 show that the proposed estimator $\bar{y}_p$ is more efficient as compared to all other considered estimators in both Populations (III and IV). The MSEs of both difference estimator $\bar{y}_D$ and Khalil estimator $\bar{y}_K$ are equivalent, but $\bar{y}_D$ is preferable over $\bar{y}_K$ because of unbiasedness.

3 Estimators under stratified random sampling

Consider a finite population of N identifiable units which are partitioned into L homogeneous subgroups called strata, such that the h^{th} strata consist of N_h units, where $h = 1, 2, \dots, L$ and $\sum_{h=1}^L N_h = N$. Let Y_h be the sensitive variable, which do not observe directly and X_h be the non-sensitive auxiliary variable which has a positive correlation with Y_h . Let $R_{x,h}$ be the ranks of the auxiliary variable X_h and S_h be a scrambling variable which is independent of Y_h and X_h . S_h has zero mean and variance $S_{S_h}^2$. The respondent is asked to give a scrambled response for the study variable Y_h given by $Z_h = Y_h + S_h$, additionally asked to provide a true response for X_h .

A simple random sample of size n_h is drawn without replacement such that $\sum_{h=1}^L n_h = n$. Let $(x_{hi}, r_{x,hi}, y_{hi}, z_{hi})$ be the observed values and $(X_{hi}, R_{x,hi}, Y_{hi}, Z_{hi})$ be the actual values on the variables $(X_h, R_{x,h}, Y_h, Z_h)$ of the i^{th} ($i = 1, 2, \dots, n$) sampled units in the h^{th} stratum. Then the measurement errors be $V_{hi}^* = x_{hi}^* - X_{hi}^*$, $U_{hi}^* = z_{hi}^* - Z_{hi}^*$ and $T_{hi} = r_{x,hi} - R_{x,hi}$. These measurement errors are assumed to be uncorrelated and having normal distribution with zero mean and variances S_{hV}^2 , S_{hU}^2 and S_{hT}^2 respectively. Let S_{hX}^2 , $S_{hR_x}^2$ and S_{hZ}^2 be the population variances; ρ_{hXZ} , ρ_{hXR_x} and ρ_{hZR_x} be the coefficients of correlation, between their subscripts.

3.1 Existing estimators in literature

In this section we consider the following existing estimators.

Mean estimator. The usual unbiased mean per unit estimator, is given by

$$\bar{y}_{S(0)} = \sum_{h=1}^L P_h \bar{z}_h, \quad (22)$$

where $P_h = \frac{N_h}{N}$ is the known stratum weight and $\bar{z}_h$ is the mean of the sensitive variable Z_h in the stratum h , (see Eq (33)). The variance of $\bar{y}_{S(0)}$, is given by

$$Var(\bar{y}_{S(0)}) = \sum_{h=1}^L P_h^2 \lambda_h (S_{hZ}^2 + S_{hU}^2), \quad (23)$$

where $\lambda_h = (n_h^{-1} - N_h^{-1})$.

Ratio estimator. The traditional ratio estimator, is given by

$$\bar{y}_{S(R)} = \sum_{h=1}^L P_h \frac{\bar{z}_h}{\bar{x}_h} \bar{X}_h, \quad (24)$$

where $\bar{X}_h$ is the known population mean and $\bar{x}_h$ is the sample mean of the auxiliary variable in stratum h , (see Eq (34)). The bias and mean square error of $\bar{y}_{S(R)}$, are given by

$$B(\bar{y}_{S(R)}) \cong \sum_{h=1}^L P_h \frac{\lambda_h}{\bar{X}_h} [R_h (S_{hX}^2 + S_{hV}^2) - \rho_{hXZ} S_{hX} S_{hZ}] \quad (25)$$

and

$$MSE(\bar{y}_{S(R)}) \cong \sum_{h=1}^L P_h^2 \lambda_h [(S_{hZ}^2 + S_{hU}^2) + R_h^2 (S_{hX}^2 + S_{hV}^2) - 2R_h \rho_{hXZ} S_{hX} S_{hZ}], \quad (26)$$

where $R_h = \frac{\bar{z}_h}{\bar{x}_h}$.

Difference estimator. The usual difference estimator, is given by

$$\bar{y}_{S(D)} = \sum_{h=1}^L P_h [\bar{z}_h + d_h (\bar{X}_h - \bar{x}_h)], \quad (27)$$

where d_h is the constant, whose value is to be determined optimally. The minimum variance of $\bar{y}_{S(D)}$, is given by

$$Var(\bar{y}_{S(D)})_{min} = \sum_{h=1}^L P_h^2 \lambda_h \left[S_{hZ}^2 + S_{hU}^2 - \frac{\rho_{hXZ}^2 S_{hZ}^2 S_{hX}^2}{(S_{hX}^2 + S_{hV}^2)} \right], \quad (28)$$

where $d_{h(opt)} = \frac{\rho_{hXZ} S_{hZ} S_{hX}}{(S_{hX}^2 + S_{hV}^2)}$.

Khalil randomized response estimator. [21] proposed the estimator, which is given by,

$$\bar{y}_{S(K)} = \sum_{h=1}^L P_h \left[(\bar{z}_h + k(\bar{X}_h - \bar{x}_h)) \left(\frac{\bar{W}_h}{\bar{w}_h} \right)^g \right], \quad (29)$$

where $\bar{w}_h = \phi_h(\alpha_h \bar{x}_h + \gamma_h) + (1 - \phi_h)(\alpha_h \bar{X}_h + \gamma_h)$ and $\bar{W}_h = \alpha_h \bar{X}_h + \gamma_h$;

k and g are constants, and ϕ_h is assumed to be an unknown constant whose value is to be determined from optimality considerations $\phi_h = \frac{\rho_{hXZ} S_{hZ} S_{hX} - k(S_{hX}^2 + S_{hV}^2)}{g R_h (S_{hX}^2 + S_{hV}^2)}$. Also $\alpha_h (\neq 0)$ and γ_h are assumed to be some known parameters of the auxiliary variable X . The bias and minimum MSE of $\bar{y}_{S(K)}$, are given by

$$B(\bar{y}_{S(K)}) \cong \sum_{h=1}^L P_h \frac{\lambda_h}{\bar{Z}_h} \left[\frac{g(g+1)}{2} \phi_h^2 R_h^2 (S_{hX}^2 + S_{hV}^2) - g \phi_h R_h \rho_{hXZ} S_{hX} S_{hZ} + g \phi_h k R (S_X^2 + S_V^2) \right] \quad (30)$$

and

$$MSE(\bar{y}_{S(K)})_{min} \cong \sum_{h=1}^L P_h^2 \lambda_h \left[S_{hZ}^2 + S_{hU}^2 - \frac{\rho_{hXZ}^2 S_{hZ}^2 S_{hX}^2}{(S_{hX}^2 + S_{hV}^2)} \right]. \quad (31)$$

which is exactly equal to the variance of the difference estimator $\bar{y}_{S(D)}$, but $\bar{y}_{S(D)}$ is preferable over $\bar{y}_{S(K)}$ because of unbiasedness.

3.2 The proposed estimator

An improved randomized response estimator for estimating the population mean of a sensitive variable in the presence of measurement error is proposed. A scrambling response of Y_h is observed in the form of $Z_h = Y_h + S_h$, where S_h is distributed as $N(0, S_{sh}^2)$. The suggested estimator is given by

$$\bar{y}_{S(P)} = \sum_{h=1}^L P_h \left[(\bar{z}_h + m_{1h}(\bar{X}_h - \bar{x}_h)) \frac{\bar{X}_h}{\bar{x}_h} + m_{2h}(\bar{R}_{x_h} - \bar{r}_{x_h}) \frac{\bar{R}_{x_h}}{\bar{r}_{x_h}} \right], \quad (32)$$

where, m_{1h} and m_{2h} are constants whose values are to be determined. $\bar{R}_{x_h}$ and $\bar{r}_{x_h}$ are the

population mean and sample mean of the ranked of the auxiliary variable, respectively (see Eq (35)). For obtaining the bias and mean square error, we define:

$$\delta_{hZ} = \sum_{i=1}^{n_h} (Z_{hi} - \bar{Z}_h), \delta_{hU} = \sum_{i=1}^{n_h} U_{hi}.$$

$$\delta_{hX} = \sum_{i=1}^{n_h} (X_{hi} - \bar{X}_h), \delta_{hV} = \sum_{i=1}^{n_h} V_{hi}.$$

$$\delta_{hR_x} = \sum_{i=1}^{n_h} (R_{x,hi} - \bar{R}_{x_h}), \delta_{hT} = \sum_{i=1}^{n_h} T_{hi}.$$

Adding δ_{hZ} and δ_{hU} , we get

$$\delta_{hZ} + \delta_{hU} = \sum_{i=1}^{n_h} (Z_{hi} - \bar{Z}_h) + \sum_{i=1}^{n_h} U_{hi}.$$

Dividing both sides by n_h , and then simplifying, we get

$$\bar{z}_h = \bar{Z}_h + \frac{1}{n_h} (\delta_{hZ} + \delta_{hU}). \quad (33)$$

Similarly, we can get

$$\bar{x}_h = \bar{X}_h + \frac{1}{n_h} (\delta_{hX} + \delta_{hV}), \quad (34)$$

and

$$\bar{r}_{x_h} = \bar{R}_{x_h} + \frac{1}{n_h} (\delta_{hR_x} + \delta_{hT}). \quad (35)$$

Let $W_{hZ} = \frac{\delta_{hZ} + \delta_{hU}}{n_h}$, $W_{hX} = \frac{\delta_{hX} + \delta_{hV}}{n_h}$ and $W_{hR_x} = \frac{\delta_{hR_x} + \delta_{hT}}{n_h}$.

In order to get the bias and MSE of the suggested estimator, we consider the following relative error terms:

$$\begin{aligned} \text{Let, } e_{0h} &= \frac{\bar{z}_h - \bar{Z}_h}{\bar{Z}_h} = \frac{W_{hZ}}{\bar{Z}_h}, e_{1h} = \frac{\bar{x}_h - \bar{X}_h}{\bar{X}_h} = \frac{W_{hX}}{\bar{X}_h}, e_{2h} = \frac{\bar{r}_{x_h} - \bar{R}_{x_h}}{\bar{R}_{x_h}} = \frac{W_{hR_x}}{\bar{R}_{x_h}}, E(e_{jh}) = 0, j = 0, 1, 2. \\ E(e_{0h}^2) &= \frac{\lambda_h(S_{hZ}^2 + S_{hU}^2)}{\bar{Z}_h^2}, E(e_{1h}^2) = \frac{\lambda_h(S_{hX}^2 + S_{hV}^2)}{\bar{X}_h^2}, E(e_{2h}^2) = \frac{\lambda_h(S_{hR_x}^2 + S_{hT}^2)}{\bar{R}_{x_h}^2}, E(e_{0h}e_{1h}) = \frac{\lambda_h \rho_{hXZ} S_{hX} S_{hZ}}{\bar{X}_h \bar{Z}_h}, \\ E(e_{0h}e_{2h}) &= \frac{\lambda_h \rho_{hR_xZ} S_{hZ} S_{hR_x}}{\bar{Z}_h \bar{R}_{x_h}} \text{ and } E(e_{1h}e_{2h}) = \frac{\lambda_h \rho_{hR_xX} S_{hR_x} S_{hX}}{\bar{X}_h \bar{R}_{x_h}}. \end{aligned}$$

Using Eq (32) in terms of errors, we have

$$\bar{y}_{S(P)} = \sum_{h=1}^L P_h \left[\bar{Z}_h (1 + e_{0h}) - m_{1h} \bar{X}_h \left(\frac{e_{1h}}{1 + e_{1h}} \right) - m_{2h} \bar{R}_{x_h} \left(\frac{e_{2h}}{1 + e_{2h}} \right) \right]. \quad (36)$$

Further simplifying, and keeping the terms up to power 2, we have

$$\bar{y}_{S(P)} - \bar{Z} \cong \sum_{h=1}^L P_h [\bar{Z}_h e_{0h} - m_{1h} \bar{X}_h e_{1h} + m_{1h} \bar{X}_h e_{1h}^2 - m_{2h} \bar{R}_{x_h} e_{2h} + m_{2h} \bar{R}_{x_h} e_{2h}^2]. \quad (37)$$

Using above equation, the bias of $\bar{y}_{S(P)}$, is given by

$$B(\bar{y}_{S(P)}) \cong \sum_{h=1}^L P_h \left[m_{1h} \frac{\lambda_h (S_{hX}^2 + S_{hV}^2)}{\bar{X}_h} + m_{2h} \frac{\lambda_h (S_{hR_x}^2 + S_{hT}^2)}{\bar{R}_{x_h}} \right]. \quad (38)$$

Squaring and then taking expectations of Eq (37), we have

$$\begin{aligned} MSE(\bar{y}_{S(P)}) \cong & \sum_{h=1}^L P_h^2 [\lambda_h (S_{hZ}^2 + S_{hU}^2) + m_{1h}^2 \lambda_h (S_{hX}^2 + S_{hV}^2) + m_{2h}^2 \lambda_h (S_{hR_x}^2 + S_{hT}^2) \\ & + 2m_{1h} m_{2h} \lambda_h \rho_{hXR_x} S_{hX} S_{hR_x} - 2m_{1h} \lambda_h \rho_{hXZ} S_{hX} S_{hZ} - 2m_{2h} \lambda_h \rho_{hZR_x} S_{hR_x} S_{hZ}]. \end{aligned} \quad (39)$$

From Eq (39), the optimum values of m_{1h} and m_{2h} are

$$m_{1h(opt)} = \frac{\lambda_h^2 (\rho_{hXR_x} S_{hX} S_{hR_x}) (\rho_{hZR_x} S_{hZ} S_{hR_x}) - \lambda_h^2 (S_{hR_x}^2 + S_{hT}^2) (\rho_{hXZ} S_{hX} S_{hZ})}{\lambda_h^2 (\rho_{hXR_x} S_{hX} S_{hR_x})^2 - \lambda_h^2 (S_{hR_x}^2 + S_{hT}^2) (S_{hX}^2 + S_{hV}^2)}, \quad (40)$$

and

$$m_{2h(opt)} = \frac{\lambda_h^2 (\rho_{hXR_x} S_{hX} S_{hR_x}) (\rho_{hXZ} S_{hX} S_{hZ}) - \lambda_h^2 (S_{hX}^2 + S_{hV}^2) (\rho_{hZR_x} S_{hZ} S_{hR_x})}{\lambda_h^2 (\rho_{hXR_x} S_{hX} S_{hR_x})^2 - \lambda_h^2 (S_{hR_x}^2 + S_{hT}^2) (S_{hX}^2 + S_{hV}^2)}, \quad (41)$$

Substitute the optimum values of m_{1h} and m_{2h} in Eq (39), the minimum MSE $\bar{y}_{S(P)}$ is given by

$$MSE(\bar{y}_{S(P)})_{min} \cong \sum_{h=1}^L P_h^2 \left[\frac{Q_h}{\lambda_h^2 (\rho_{hXR_x} S_{hX} S_{hR_x})^2 - \lambda_h^2 (S_{hR_x}^2 + S_{hT}^2) (S_{hX}^2 + S_{hV}^2)} \right]. \quad (42)$$

where,

$$\begin{aligned} Q_h = & \lambda_h^3 (S_{hZ}^2 + S_{hU}^2) (\rho_{hXR_x} S_{hX} S_{hR_x})^2 + \lambda_h^3 (S_{hR_x}^2 + S_{hT}^2) (\rho_{hXZ} S_{hX} S_{hZ})^2 \\ & + \lambda_h^3 (S_{hX}^2 + S_{hV}^2) (\rho_{hZR_x} S_{hZ} S_{hR_x})^2 - \lambda_h^3 (S_{hZ}^2 + S_{hU}^2) (S_{hX}^2 + S_{hV}^2) (S_{hR_x}^2 + S_{hT}^2) \\ & - 2\lambda_h^3 (\rho_{hXZ} S_{hX} S_{hZ}) (\rho_{hXR_x} S_{hX} S_{hR_x}) (\rho_{hZR_x} S_{hZ} S_{hR_x}). \end{aligned}$$

3.3 Efficiency comparison

The efficiency comparison of $\bar{y}_{S(0)}$, $\bar{y}_{S(R)}$, $\bar{y}_{S(D)}$ and $\bar{y}_{S(K)}$ with respect to $\bar{y}_{S(P)}$ are given by,

1. From Eqs (23) and (42)

$MSE(\bar{y}_{S(P)})_{min} < Var(\bar{y}_{S(0)})$, if

$$\sum_{h=1}^L P_h^2 \left[\frac{Q_h}{\lambda_h^2 (\rho_{hXR_x} S_{hX} S_{hR_x})^2 - \lambda_h^2 (S_{hR_x}^2 + S_{hT}^2) (S_{hX}^2 + S_{hV}^2)} \right] - \sum_{h=1}^L P_h^2 \lambda_h (S_{hZ}^2 + S_{hU}^2) < 0$$

2. From Eqs (26) and (42)

$$\begin{aligned} MSE(\bar{y}_{S(P)})_{min} < MSE(\bar{y}_{S(R)}), \text{ if } & \sum_{h=1}^L P_h^2 \left[\frac{Q_h}{\lambda_h^2 (\rho_{hXR_x} S_{hX} S_{hR_x})^2 - \lambda_h^2 (S_{hR_x}^2 + S_{hT}^2) (S_{hX}^2 + S_{hV}^2)} \right] - \\ & \sum_{h=1}^L P_h^2 \lambda_h [(S_{hZ}^2 + S_{hU}^2) + R_h^2 (S_{hX}^2 + S_{hV}^2) - 2R_h \rho_{hXZ} S_{hX} S_{hZ}] < 0 \end{aligned}$$

3. From Eqs (28) and (42)

$MSE(\bar{y}_{S(P)})_{min} < Var(\bar{y}_{S(D)})$, if

$$\sum_{h=1}^L P_h^2 \left[\frac{Q_h}{\lambda_h^2 (\rho_{hXR_x} S_{hX} S_{hR_x})^2 - \lambda_h^2 (S_{hR_x}^2 + S_{hT}^2) (S_{hX}^2 + S_{hV}^2)} \right] - \sum_{h=1}^L P_h^2 \lambda_h \left[S_{hZ}^2 + S_{hU}^2 - \frac{\rho_{hXZ}^2 S_{hX}^2 S_{hZ}^2}{(S_{hX}^2 + S_{hV}^2)} \right] < 0$$

Table 7. MSE of different estimators for Population V under simulation.

Estimators	MSE without measurement error			MSE with measurement error		
	n_h			n_h		
	10%	15%	20%	10%	15%	20%
$\bar{y}_{S(0)}$	38.3442	17.041870	9.941091	38.366670	17.052020	9.947028
$\bar{y}_{S(R)}$	9.801616	4.356274	2.541160	9.832481	4.370093	2.549331
$\bar{y}_{S(D)}, \bar{y}_{S(K)}$	9.675700	4.300311	2.508515	9.705609	4.313713	2.516432
$\bar{y}_{S(P)}$	8.422953	3.743535	2.183729	8.445560	3.753748	2.189704

<https://doi.org/10.1371/journal.pone.0212111.t007>

4. From Eqs (31) and (42)

$MSE(\bar{y}_{S(P)})_{min} < MSE(\bar{y}_{S(K)})$, if

$$\sum_{h=1}^L P_h^2 \left[\frac{Q_h}{\lambda_h^2 (\rho_{hX R_X} S_{hX} S_{hR_X})^2 - \lambda_h^2 (S_{hR_X}^2 + S_{hT}^2)(S_{hX}^2 + S_{hV}^2)} \right] - \sum_{h=1}^L P_h^2 \lambda_h \left[S_{hZ}^2 + S_{hU}^2 - \frac{\rho_{hXZ}^2 S_{hZ}^2 S_{hX}^2}{(S_{hX}^2 + S_{hV}^2)} \right] < 0$$

The proposed class of estimators is more efficient than other existing estimators when above Conditions 1 to 4 are satisfied.

3.4 Numerical results

In this section two populations are generated for simulation study and one for real data set.

3.4.1 Simulation study. We have generated two populations of size 1,000 from multivariate normal distribution with different covariance matrices. The results are given in Tables 7 and 8. The mean and covariance matrices are give below

Population V.

$$\mu_5 = \begin{bmatrix} 250 \\ 250 \\ 250 \end{bmatrix}$$

and

$$\sum_5 = \begin{bmatrix} 1000 & 800 & 810 \\ 800 & 850 & 820 \\ 810 & 820 & 840 \end{bmatrix}$$

Table 8. MSE of different estimators for Population VI under simulation.

Estimators	MSE without measurement error			MSE with measurement error		
	n_h			n_h		
	10%	15%	20%	10%	15%	20%
$\bar{y}_{S(0)}$	25.959630	11.537613	6.730274	25.982093	11.547763	6.736212
$\bar{y}_{S(R)}$	12.978994	5.768442	3.364924	13.009863	5.782262	3.373097
$\bar{y}_{S(D)}, \bar{y}_{S(K)}$	11.890829	5.284813	3.082808	11.918461	5.297220	3.090119
$\bar{y}_{S(P)}$	9.917212	4.407650	2.571129	9.941450	4.418575	2.577538

<https://doi.org/10.1371/journal.pone.0212111.t008>

$$N_1 = 500 \text{ and } N_2 = 500,$$

$$\rho_{1XY} = 0.8554, \rho_{1XR_x} = 0.9231 \text{ and } \rho_{1YR_x} = 0.8705$$

$$\rho_{2XY} = 0.8797, \rho_{2XR_x} = 0.9725 \text{ and } \rho_{2YR_x} = 0.8940$$

Population VI.

$$\mu_6 = \begin{bmatrix} 250 \\ 250 \\ 250 \end{bmatrix}$$

and

$$\sum_6 = \begin{bmatrix} 675 & 470 & 420 \\ 470 & 600 & 400 \\ 420 & 400 & 500 \end{bmatrix}$$

$$N_1 = 400 \text{ and } N_2 = 600$$

$$\rho_{1XY} = 0.7172, \rho_{1XR_x} = 0.7321 \text{ and } \rho_{1YR_x} = 0.7104$$

$$\rho_{2XY} = 0.7592, \rho_{2XR_x} = 0.7496 \text{ and } \rho_{2YR_x} = 0.7492$$

Covariance matrices show the distribution of sensitive variable Y_h , the auxiliary variable X_h and the ranks of the auxiliary variable $R_{x,h}$. Population V consist of two equal strata and Population VI comprises of two unequal strata. In Population V there is high correlation among the variables, and low correlation in Population VI. The scrambling response S_h is distributed as $N(0, 0.01\sigma_{X_h})$. The response variable is $Z_h = Y_h + S_h$. We estimate the MSE using $k_h = 1000$ samples of various sizes selected from each strata. Three different sample sizes, 10%, 15% and 20% are taken for both populations. The expression is given below:

$$MSE(\bar{y}_{S(i)}) = \frac{1}{k_h} \sum_{j=1}^{k_h} (y_{hij} - \bar{y}_h)^2.$$

where $i = 0, R, D, K, P$

Tables 7 and 8 show that the estimator $\bar{y}_{S(P)}$ performs better as compared to the estimators $\bar{y}_{S(0)}, \bar{y}_{S(R)}, \bar{y}_{S(D)}$ and $\bar{y}_{S(K)}$. The efficiency of the estimator $\bar{y}_{S(P)}$ is improved when there is sufficient correlation between the study variable and the auxiliary variable. By increasing the sample size, MSE values decreases. As the MSEs of $\bar{y}_{S(D)}$ and $\bar{y}_{S(K)}$ are equal, so their numerical results are also identical for both the populations.

3.4.2 Application to real data. In this section we consider the real life data set for numerical comparisons. Strata I consist of 318 observations and Strata II contain 336 observations. The data summary is given below (see Tables 9 and 10). The results are given in Table 11.

$$\rho_{1XY} = 0.7564, \rho_{1XR_x} = 0.7831 \text{ and } \rho_{1YR_x} = 0.6151$$

Table 9. Data summary strata I.

Variable	Mean	st.Dev	Min	Med	Max	C.V
Forced expiratory volume (Y_1)	2.45	0.65	0.79	2.48	3.83	0.26
Age (X_1)	9.84	2.93	3.00	10.00	19.00	0.29
Smoke (S_1) 0,1	0.12	0.32	0.00	0.00	1.00	2.66

<https://doi.org/10.1371/journal.pone.0212111.t009>

Table 10. Data summary strata II.

Variable	Mean	st.Dev	Min	Med	Max	C.V
Forced expiratory volume (Y_2)	2.68	1.00	0.79	2.61	5.79	0.37
Age (X_2)	10.01	2.97	3.00	10.00	19.00	0.29
Sex (S_2)0,1	0.07	0.27	0.00	0.00	1.00	3.85

<https://doi.org/10.1371/journal.pone.0212111.t010>

Table 11. MSE of different estimators for Population VII under real data.

Estimators	MSE without measurement error			MSE with measurement error		
	n_h			n_h		
	10%	15%	20%	10%	15%	20%
$\bar{y}_{S(0)}$	0.041509	0.027650	0.019442	0.083287	0.055736	0.040301
$\bar{y}_{S(R)}$	0.015310	0.010116	0.007109	0.058106	0.038821	0.028385
$\bar{y}_{S(D)}, \bar{y}_K$	0.014450	0.009574	0.006734	0.057221	0.038270	0.027991
$\bar{y}_{S(P)}$	0.014306	0.009480	0.006667	0.056736	0.037970	0.027792

<https://doi.org/10.1371/journal.pone.0212111.t011>

$$\rho_{2XY} = 0.8109, \rho_{2XR_x} = 0.7765 \text{ and } \rho_{2YR_x} = 0.6575$$

Population VII. (Source: [24])

In Table 11, we observed that the estimator $\bar{y}_{S(P)}$ performs better than the estimators $\bar{y}_{S(0)}, \bar{y}_{S(R)}, \bar{y}_{S(D)}$ and $\bar{y}_{S(K)}$. The estimators $\bar{y}_{S(D)}$ and $\bar{y}_{S(K)}$ have same MSEs but $\bar{y}_{S(D)}$ is preferable due to unbiasedness. As the sample size increases the MSE values decreases, which are the expected results.

4 Conclusion

In the present paper, we have proposed a new improved estimator of the finite population mean that encounter additional information on the auxiliary variable as well as on ranks of the auxiliary variable in the presence of measurement error under simple and stratified random sampling. Through simulation study and real life data sets (see Tables 1, 2, 5, 6, 7, 8 and 11) it is observed that the proposed estimators $\bar{y}_p$ and $\bar{y}_{S(P)}$ perform better than the existing estimators, particularly when there is sufficient correlation between the study variable and the auxiliary variable. It is also concluded that difference estimator and [21] estimator are equally efficient, but difference estimator is preferable due to unbiasedness.

Supporting information

S1 File. Data used in the manuscript “S1_File.csv”.
(CSV)

Acknowledgments

The authors are grateful to anonymous referees for their valuable comments and feedback.

Author Contributions

Data curation: Erum Zahid.

Formal analysis: Erum Zahid.

Funding acquisition: Erum Zahid.

Investigation: Javid Shabbir.

Methodology: Erum Zahid.

Project administration: Javid Shabbir.

Software: Erum Zahid.

Supervision: Javid Shabbir.

Validation: Erum Zahid.

Visualization: Erum Zahid, Javid Shabbir.

Writing – original draft: Erum Zahid.

Writing – review & editing: Erum Zahid, Javid Shabbir.

References

1. Cochran WG. Errors of measurement in statistics. *Technometrics*. 1968; 10(4):637–666. <https://doi.org/10.2307/1267450>
2. Fuller WA. Estimation in the presence of measurement error. *International Statistical Review/Revue Internationale de Statistique*. 1995; p. 121–141.
3. Shalabh S. Ratio method of estimation in the presence of measurement errors. *Jour Ind Soc Agri Statist*. 1997; 52:150–155.
4. Biemer PP, Groves RM, Lyberg LE, Mathiowetz NA, Sudman S. *Measurement Errors in Surveys*. John Wiley & Sons; 2011.
5. Shukla D, Pathak S, Thakur N. An estimator for mean estimation in presence of measurement error. *Research and Reviews: A Journal of Statistics*. 2012; 1(1):1–8.
6. Singh RS, Sharma P. Method of Estimation in the Presence of Non-response and Measurement Errors Simultaneously. *Journal of Modern Applied Statistical Methods*. 2015; 14(1):12. <https://doi.org/10.22237/jmasm/1430453460>
7. Kumar S. Improved estimation of population mean in presence of nonresponse and measurement error. *Journal of Statistical Theory and Practice*. 2016; 10(4):707–720. <https://doi.org/10.1080/15598608.2016.1216488>
8. Azeem M, Hanif M. Joint influence of measurement error and non response on estimation of population mean. *Communications in Statistics-Theory and Methods*. 2017; 46(4):1679–1693. <https://doi.org/10.1080/03610926.2015.1026992>
9. Zahid E, Shabbir J. Estimation of population mean in the presence of measurement error and non response under stratified random sampling. *PloS one*. 2018; 13(2):e0191572. <https://doi.org/10.1371/journal.pone.0191572> PMID: 29401519
10. Warner SL. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*. 1965; 60(309):63–69. <https://doi.org/10.1080/01621459.1965.10480775> PMID: 12261830
11. Greenberg BG, Kuebler RR Jr, Abernathy JR, Horvitz DG. Application of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association*. 1971; 66(334):243–250. <https://doi.org/10.1080/01621459.1971.10482248>
12. Eichhorn BH, Hayre LS. Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and inference*. 1983; 7(4):307–316. [https://doi.org/10.1016/0378-3758\(83\)90002-2](https://doi.org/10.1016/0378-3758(83)90002-2)
13. Gupta S, Shabbir J. Sensitivity estimation for personal interview survey questions. *Statistica*. 2004; 64(4):643–653.
14. Gupta S, Shabbir J, Sehra S. Mean and sensitivity estimation in optional randomized response models. *Journal of Statistical Planning and Inference*. 2010; 140(10):2870–2874. <https://doi.org/10.1016/j.jspi.2010.03.010>
15. Singh HP, Mathur N. Estimation of population mean when coefficient of variation is known using scrambled response technique. *Journal of statistical planning and inference*. 2005; 131(1):135–144. <https://doi.org/10.1016/j.jspi.2004.01.002>

16. Chaudhuri A, Pal S. On efficacy of empirical Bayes estimation of a finite population mean of a sensitive variable through randomized responses. *Model Assisted Statistics and Applications*. 2015; 10(4):283–288. <https://doi.org/10.3233/MAS-150332>
17. Gupta S, Shabbir J, Sousa R, Corte-Real P. Improved Exponential Type Estimators of the Mean of a Sensitive Variable in the Presence of Nonsensitive Auxiliary Information. *Communications in Statistics-Simulation and Computation*. 2016; 45(9):3317–3328. <https://doi.org/10.1080/03610918.2014.941487>
18. Shabbir J, Gupta S. On estimating finite population mean in simple and stratified random sampling. *Communications in Statistics-Theory and Methods*. 2010; 40(2):199–212. <https://doi.org/10.1080/03610920903411259>
19. Haq A, Shabbir J. Improved family of ratio estimators in simple and stratified random sampling. *Communications in Statistics-Theory and Methods*. 2013; 42(5):782–799. <https://doi.org/10.1080/03610926.2011.579377>
20. Haq A, Khan M, Hussain Z. A new estimator of finite population mean based on the dual use of the auxiliary information. *Communications in Statistics-Theory and Methods*. 2017; 46(9):4425–4436. <https://doi.org/10.1080/03610926.2015.1083112>
21. Khalil S, Gupta S, Hanif M. Estimation of finite population mean in stratified sampling using scrambled responses in the presence of measurement errors. *Communications in Statistics-Theory and Methods*. 2018; p. 1–9. <https://doi.org/10.1080/03610926.2018.1435817>
22. Cochran WG. *Sampling Techniques*: 3d Ed. Wiley New York; 1977.
23. Sukhatme PV, Sukhatme B, Sukhatme S, Asok C. *Sampling theory with applications*. Indian Society of Agricultural Statistics, New Delhi & IOWA State University Press, Ames, USA. 1984;.
24. Rosner B. *Fundamentals of Biostatistics*. 2006. Duxbury Press. 2015;.