International Journal of Statistical Sciences *ISSN 1683-5603*
Vol.24(1), March, 2024, pp 155-170
DOI: https://doi.org/10.3329/ijss.v24i1.72035

Indirect Questioning Technique Related to Sensitive Quantitative Variables with Options for Direct, Randomized and Item Count Responses

Purnima Shaw[1*] and Sanghamitra Pal[2]

[1]Department of Statistics and Information Management, Reserve Bank of India, New Delhi-110001, India.

[2]Department of Statistics, West Bengal State University, West Bengal-700126, India.
Email: mitrapal2013@gmail.com

[*]Correspondence should be addressed to Purnima Shaw
(Email: purnimashaw2011@gmail.com)

The views expressed in this paper are personal and not of the Reserve Bank of India.

[Received November 15, 2023; Accepted February 8, 2024]

Abstract

Randomized Response (RR) Technique (RRT) pioneered by Warner (1965) is a useful tool to elicit responses on sensitive characteristics, such as induced abortions, drug abuse, drunken driving, total amount of counterfeit notes of a particular denomination held by individuals in the population*, etc*. There exists a huge literature on Randomized Response (RR) devices for estimation of finite population mean of quantitative variables, sensitive in nature mostly based on Eichhorn and Hayre (1983). Device-I and Device-II *vide* Chaudhuri and Christofides (2013) allow estimation of population mean of sensitive quantitative variables using sample chosen by a general sampling design. On the other hand, Item Count Technique (ICT), described elaborately in Chaudhuri and Christofides (2013), is an alternative to RRT for respondents who do not choose to provide RRs. While some respondents may find a variable as sensitive, others may find it innocuous enough to provide a direct response (DR) about his/her true value. In such a case, Optional Randomized Response (ORR) Technique (ORRT) with options for DR and RR was introduced by Chaudhuri and Mukherjee (1985). Pal (2007) proposed an ORR device which offers choices for RR and ICT to the respondents for giving their answers. A new ORRT with options for DR, RR and ICT was provided by Shaw and Pal (2021) for eliciting indirect responses on sensitive characteristics. As this device relates to estimation of population proportion of sensitive characteristics, an attempt has been made to extend it for sensitive quantitative variables. Further, to take care of individuals' varying choices for DR, RR and ICT and to protect the privacy of the respondents' choices, this paper develops an ORR device allowing the respondents chosen by a general sampling design, to choose any one of the three options according to their choices.

Keywords and Phrases: Item Count Technique; Optional Randomized Response Technique; Quantitative Randomized Response; Stigmatizing Characteristics; Varying Probability Sampling.

AMS Classification: 62DO5.

1. Introduction

Warner's (1965) pioneering RR device is well-known for estimating finite population proportion of stigmatizing characteristics. Given the importance of estimating finite population mean or total of a sensitive quantitative variable, Greenberg, Kuebler, Abernathy and Horvitz (1971), Fox and Tracy (1984) and Eichhorn and Hayre (1983) among others, are important contributors of RRT's. The initial literatures which allowed drawing of sample by a general sampling scheme include Eriksson (1973), Adhikari, Chaudhuri and Vijayan (1984), Chaudhuri (1987), Sengupta and Kundu (1989) and Chaudhuri (1992). Subsequently, Chaudhuri (2011) and Chaudhuri and Christofides (2013) presented two RR devices *viz.*, Device-I and Device-II which are simple to execute and samples can be drawn using a general sampling design. Consider a finite population $U = (1,2, \dots, i, \dots, N)$ consisting of a known number N of individuals. Let y_i denote the value a stigmatizing quantitative variable for the i^{th}, $(i = 1,2, \dots, N)$ individual in the population. The objective is to estimate the population mean of the y-values, *viz.*,

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{1}$$

In Device-I, an individual i is requested to draw one card independently from each of the two boxes provided, wherein the first box contains identical cards bearing some numbers and the second box contains identical cards bearing numbers different from those in the first box. The individual is asked to report his/her value $a_j y_i + b_k$, where a_j and b_k are the numbers in the cards drawn from the first and second box, respectively. On the other hand, in Device-II, a box containing identical cards, out of which a proportion $K(0 < K < 1)$ are marked "True" and the rest of the cards bearing real numbers in different proportions, is provided to the respondents. A respondent i is requested to select one of the cards and report the true value y_i if the card bears "True", else report the value written in the card. Using these two devices, given a sample selected by an unequal probability sampling design, y_i can be estimated for each i, followed by estimation of $\bar{Y}$ and its variance.

In order to overcome the respondents' fear of revelation of privacy in the RR devices, Item Count Technique, also known as the Block Total Response or the Unmatched Count Technique, was introduced by Raghavarao and Federer (1979), Miller (1984) and Miller *et al*. (1986). An important development in this area includes Chaudhuri and Christofides (2007). For estimating finite population mean of a sensitive quantitative variable, Chaudhuri and Christofides (2013) presented an ICT, in which two independent samples are required to be drawn from the population by using a general sampling scheme. An individual in the first sample is given a questionnaire containing $(G + 1)$ quantitative items, where G items are innocuous and the $(G + 1)^{th}$ item is related to the sensitive variable. The individual is requested to answer the sum of the $(G + 1)$ items without revealing the individual values. An individual in the second sample is provided with a questionnaire containing the same G innocuous items as in the questionnaire used for the first sample. The person is requested to report the sum of these G items without revealing the responses for the individual items. This procedure was later improved upon by Shaw (2015). The literature on a similar procedure named as Item Sum Technique includes Trappmann *et al.* (2014).

While some individuals find a variable sensitive and choose to mask their responses with the help of a RR device, others may find it innocuous and thus may not hesitate to provide a DR. To take care of such circumstances, Chaudhuri and Christofides (2013) present ORR devices extensively, including those by Arnab (2004), Gupta *et al.* (2010) and Huang (2010). Chaudhuri (2011) extended the Chaudhuri and Dihidar's (2009) ORRT (for estimating population proportion of individuals bearing a sensitive characteristic) to cover quantitative variable, sensitive in nature. In this ORR device, a sampled respondent i may choose to answer the true value y_i directly with an unknown probability C_i, $(0 \leq C_i \leq 1)$, or may opt to give an RR with probability $(1 - C_i)$, without revealing the choice. Pal (2008) also presents an important ORR device providing options for DR and RR. Arnab and Rueda (2016) provide an elaborate discussion on the ORR devices existing in the literature. Recently, Arnab (2018) and Pal *et al*. (2020) presented important contributions to the ORR techniques.

Pal (2007) presented an ORR device for estimating the population proportion of a sensitive characteristic, say A, in which the respondents are provided options for RR and ICT. This method also requires selection of two independent samples from the population. An individual in the first sample or second sample, opting for RR has to choose a random number from $(0,1,2,\dots,G)$ and report the number which he/she gets after adding it with the true value (1, if the individual bears the sensitive characteristic and 0, otherwise). Respondents in the first sample opting for ICT, are requested to report the total number of items statements holding true for him/her from a questionnaire containing $(G+1)$ item statements with G innocuous item statements and the $(G+1)^{th}$ item statement being "I bear characteristic A or F", where F is an innocuous characteristic unrelated to A. Respondents in the second sample opting for ICT, are requested to report the total number of statements holding true from a questionnaire containing $(G+1)$ item statements with G item statements being innocuous and same as those in the first questionnaire and the $(G+1)^{th}$ item statement is "I do not have characteristic A or I do not have characteristic F".
This paper, extends the Pal (2007) device for estimating population mean of a sensitive quantitative variable (Section 2). In practical situation, a population may contain a few individuals who choose DR, a few persons may opt RR and the rest may choose ICT. However, the ORR devices existing in the literature provides only two types of response options to the respondents *i.e.*, DR and RR. Hence, to avoid possible non-responses from potential respondents, Section 3 presents an ORR device providing all the three types of response options *i.e.*, DR, RR and ICT, using three independent samples (each sample of the same size or of different size) drawn from the population using a general sampling design from the population. To examine the performance of the two proposed devices, a numerical exercise based on simulated data is presented in Section 4. The concluding remarks are provided in Section 5.

2. Proposed ORR Device with options for RR and ICT using two Independent Samples

The device by Pal (2007), providing options for RR and ICT, covers qualitative sensitive characteristics. An ORR device with the same two response options, has been proposed in this section to cover quantitative sensitive variables for estimating $\bar{Y}$. Consider a respondent i, in the population, bears an unknown probability $C_i (0 \leq C_i \leq 1)$ of choosing RR, and with the remaining probability $(1 - C_i)$ for choosing ICT. The probability C_i is assumed to be unknown and different for each individual in the population. Consider a sample s_1 is selected from U according to an unequal probability sampling design P admitting positive first order and second order inclusion probabilities $\pi_i = \sum_{s_1 \ni i} P(s_1)$, $\pi_{ij} = \sum_{s_1 \ni i,j} P(s_1)$, $i \neq j, (i,j = 1,2,\dots,N)$. A respondent i is provided with options to either answer as per an RR device or answer according to an ICT questionnaire, without divulging the choice. If a respondent *i* opts to give RR, then, he/ she is requested to multiply the true value y_i with a number, say a_{11i}, randomly chosen from $(d_1, d_2, d_3, \dots, d_T)$, with mean $\mu_a = \frac{1}{T}\sum_{t=1}^{T} d_t = 1$, add with a number, say b_{11i}, randomly chosen from $(f_1, f_2, f_3, \dots, f_M)$, with mean $\mu_b = \frac{1}{M}\sum_{m=1}^{M} f_m$, and then report the resulting number. The numbers $(d_1, d_2, d_3, \dots, d_T)$ and $(f_1, f_2, f_3, \dots, f_M)$ provided to the respondents are decided at the discretion of the investigator. For respondents opting RR, Device-I *vide* Chaudhuri (2011) and Chaudhuri and Christofides (2013) is used for the simplicity of its execution. The questionnaire for ICT consists of G quantitative innocuous quantitative items, the $(G+1)^{th}$ quantitative item being on the value of y_i. If the respondent opts for ICT questionnaire, then he/she is requested to add the answers of all the $(G+1)$ items and respond the total, say t_{1i} to the investigator, without divulging the values of the individual items. The response for ICT, t_{1i} can be expressed as,

$$t_{1i} = \sum_{h=1}^{G} u_{ih} + y_i \tag{2}$$

with $\sum_{h=1}^{G} u_{ih}$ being the sum of the answers to the G innocuous questions in the ICT questionnaire. Consider, the i^{th} respondent's answer as z_{11i},

$$z_{11i} = \begin{cases} (a_{11i}y_i + b_{11i}) \text{ with probability } C_i \\ t_{1i} \text{ with probability } (1-C_i) \end{cases}, 0 \le C_i \le 1 \tag{3}$$

The respondent i is requested to provide another response, say z_{12i}, independent from z_{11i}, following the same procedure as for z_{11i}. Let a_{12i} be the number chosen by the respondent from $(d_1, d_2, d_3, \dots, d_T)$ and b_{12i} be the number chosen from $(f_1, f_2, f_3, \dots, f_M)$ opting RR, independent of the selection of a_{11i} and b_{11i} respectively. Then,

$$z_{12i} = \begin{cases} (a_{12i}y_i + b_{12i}) \text{ with probability } C_i \\ t_{1i} \text{ with probability } (1-C_i) \end{cases}, 0 \le C_i \le 1 \tag{4}$$

Taking E_R and V_R as the RR-based expectation and variance operators, respectively,

$$E_R(z_{11i}) = E_R(z_{12i}) = C_i y_i + \mu_b C_i + (1-C_i)\left(\sum_{h=1}^{G} u_{ih} + y_i\right), 0 \le C_i \le 1 \tag{5}$$

Considering the unbiased estimators used by Pal (2007),

let, $$r_{1i} = \frac{z_{11i} + z_{12i}}{2}, \quad v_{1i} = \frac{(z_{11i} - z_{12i})^2}{4} \tag{6}$$

Then, $$E_R(r_{1i}) = E_R(z_{11i}) = E_R(z_{12i}) \tag{7}$$

and, $$E_R(v_{1i}) = V_R(r_{1i}) \tag{8}$$

Consider a second sample s_2, selected from U (independent of the selection of s_1) according to an unequal probability sampling design P admitting positive first order and second order inclusion-probabilities $\pi_k = \sum_{s_2 \ni k} P(s_2), \pi_{kl} = \sum_{s_2 \ni k,l} P(s_2), k \neq l, (k, l = 1,2,3, \dots, N)$. A respondent k is provided with options to either answer as per an RR device or answer according to an ICT questionnaire, without divulging the chosen option. Here, the instructions for RR and ICT are different from those in the first sample. If a respondent chooses to give RR, then, he/she has to report a number, say b_{21k}, randomly chosen from $(f_1, f_2, f_3, \dots, f_M)$. The ICT questionnaire contains the same G quantitative innocuous items as used in the questionnaire for the first sample. In case ICT is chosen, then, sum of the G questions, say t_{2k} has to answered, where,

$$t_{2k} = \sum_{h=1}^{G} u_{kh} \tag{9}$$

where, $\sum_{h=1}^{G} u_{kh}$ is the sum of all the G answers for the questionnaire. Let z_{21k} be the answer from the k^{th} respondent, where,

$$z_{21k} = \begin{cases} b_{21k} \; with \; probability \; C_k \\ t_{2k} \; with \; probability \; (1 - C_k) \end{cases}, 0 \leq C_k \leq 1 \tag{10}$$

The respondent k is requested to provide another response, say z_{22k}, independent from z_{21k}, following the steps similar to above. Assuming, b_{22k} as the random number chosen from $(f_1, f_2, f_3, \dots, f_M)$, independent of b_{21k},

$$z_{22k} = \begin{cases} b_{22k} \; with \; probability \; C_k \\ t_{2k} \; with \; probability \; (1 - C_k) \end{cases}, 0 \leq C_k \leq 1 \tag{11}$$

Then, following (6), consider,

$$r_{2k} = \frac{z_{21k} + z_{22k}}{2}, \quad v_{2k} = \frac{(z_{21k} - z_{22k})^2}{4} \tag{12}$$

Then,

$$E_R(r_{2k}) = E_R(z_{21k}) = E_R(z_{22k}) \tag{13}$$

and,

$$E_R(v_{2k}) = V_R(r_{2k}) \tag{14}$$

Consider the estimator e, where,

$$e = \frac{1}{N} \sum_{i \in s_1} \frac{r_{1i}}{\pi_i} - \frac{1}{N} \sum_{k \in s_2} \frac{r_{2k}}{\pi_k} \tag{15}$$

Then, assuming E_P and V_P as the design-based expectation and variance operators, respectively,

$$E(e) = E_R E_P(e) = E_P E_R(e)$$

$$= E_P \left\{ \frac{1}{N} \sum_{i \in s_1} \frac{C_i y_i + \mu_b C_i + (1 - C_i)(\sum_{h=1}^{G} u_{ih} + y_i)}{\pi_i} \right\} - E_P \left\{ \frac{1}{N} \sum_{k \in s_2} \frac{\mu_b C_k + (1 - C_k) \sum_{h=1}^{G} u_{kh}}{\pi_k} \right\}$$

$$= \frac{1}{N}\sum_{i=1}^{N} C_i y_i + \frac{\mu_b}{N}\sum_{i=1}^{N} C_i + \frac{1}{N}\sum_{i=1}^{N} y_i - \frac{1}{N}\sum_{i=1}^{N} C_i y_i + \frac{1}{N}\sum_{i=1}^{N}(1 - C_i)\sum_{h=1}^{G} u_{ih} - \frac{\mu_b}{N}\sum_{k=1}^{N} C_k$$

$$-\frac{1}{N}\sum_{k=1}^{N}(1 - C_k)\sum_{h=1}^{G} u_{kh} = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{Y} \tag{16}$$

Hence, e is an unbiased estimator of $\bar{Y}$. Now, taking clue from Chaudhuri and Pal (2002), variance of e can be expressed as,

$$V(e) = V\left(\frac{1}{N}\sum_{i\in s_1}\frac{r_{1i}}{\pi_i}\right) + V\left(\frac{1}{N}\sum_{k\in s_2}\frac{r_{2k}}{\pi_k}\right)$$

$$= E_P V_R\left(\frac{1}{N}\sum_{i\in s_1}\frac{r_{1i}}{\pi_i}\right) + V_P E_R\left(\frac{1}{N}\sum_{i\in s_1}\frac{r_{1i}}{\pi_i}\right) + E_P V_R\left(\frac{1}{N}\sum_{k\in s_2}\frac{r_{2k}}{\pi_k}\right) + V_P E_R\left(\frac{1}{N}\sum_{k\in s_2}\frac{r_{2k}}{\pi_k}\right)$$

$$= E_R V_P\left(\frac{1}{N}\sum_{i\in s_1}\frac{r_{1i}}{\pi_i}\right) + V_R E_P\left(\frac{1}{N}\sum_{i\in s_1}\frac{r_{1i}}{\pi_i}\right) + E_R V_P\left(\frac{1}{N}\sum_{k\in s_2}\frac{r_{2k}}{\pi_k}\right) + V_R E_P\left(\frac{1}{N}\sum_{k\in s_2}\frac{r_{2k}}{\pi_k}\right)$$

$$= E_R\left[\frac{1}{N^2}\left\{\sum_{i}^{N}\sum_{<j}^{N}\left(\pi_i\pi_j - \pi_{ij}\right)\left(\frac{r_{1i}}{\pi_i} - \frac{r_{1j}}{\pi_j}\right)^2 + \sum_{i=1}^{N}\frac{\beta_i}{\pi_i}{r_{1i}}^2\right\}\right] + \frac{1}{N^2}\sum_{i=1}^{N} V_R(r_{1i})$$

$$+E_R\left[\frac{1}{N^2}\left\{\sum_{k}^{N}\sum_{<l}^{N}\left(\pi_k\pi_l - \pi_{kl}\right)\left(\frac{r_{2k}}{\pi_k} - \frac{r_{2l}}{\pi_l}\right)^2 + \sum_{k=1}^{N}\frac{\beta_k}{\pi_k}{r_{2k}}^2\right\}\right] + \frac{1}{N^2}\sum_{k=1}^{N} V_R(r_{2k}) \tag{17}$$

where, $$\beta_i = 1 + \frac{1}{\pi_i}\sum_{j\neq i}^{N}\pi_{ij} - \sum_{i=1}^{N}\pi_i, \qquad \beta_k = 1 + \frac{1}{\pi_k}\sum_{k\neq l}^{N}\pi_{kl} - \sum_{k=1}^{N}\pi_k$$

If every sample s_1 and s_2 contains common number of distinct units in it, then, $\beta_i = 0 \; \forall$ i and $\beta_k = 0 \; \forall$ k throughout in $V(e)$ above. Then, taking clue from Chaudhuri and Pal (2002), an unbiased estimator for $V(e)$ is,

$$v(e) = \frac{1}{N^2}\left\{\sum_{i<}\sum_{j\in s_1}\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{r_{1i}}{\pi_i} - \frac{r_{1j}}{\pi_j}\right)^2 + \sum_{i\in s_1}\frac{\beta_i}{{\pi_i}^2}{r_{1i}}^2\right\} + \frac{1}{N^2}\sum_{i\in s_1}\frac{v_{1i}}{\pi_i} \tag{18}$$

$$+\frac{1}{N^2}\left\{\sum_{k<}\sum_{l\in s_2}\left(\frac{\pi_k\pi_l-\pi_{kl}}{\pi_{kl}}\right)\left(\frac{r_{2k}}{\pi_k}-\frac{r_{2l}}{\pi_l}\right)^2+\sum_{k\in s_2}\frac{\beta_k}{{\pi_k}^2}{r_{2k}}^2\right\}+\frac{1}{N^2}\sum_{k\in s_2}\frac{v_{2k}}{\pi_k}$$

with $\beta_i = 0 \; \forall$ i and $\beta_k = 0 \; \forall$ k in $v(e)$ when applicable. Hence, $v(e)$ is an unbiased estimator of $V(e)$, such that $E\{v(e)\} = E_P E_R\{v(e)\} = E_R E_P\{v(e)\} = V(e)$. A $100(1-\alpha)\%$ Confidence Interval for $\bar{Y}$ is, $[L, U]$, where,

$$L = e - \left(\tau_{\alpha/2}\sqrt{v(e)}\right), \qquad U = e + \left(\tau_{\alpha/2}\sqrt{v(e)}\right) \tag{19}$$

$\tau_{\alpha/2}$ is the upper $\frac{\alpha}{2}$ point of Standard Normal distribution

3. Proposed ORR Device with Options for DR, RR and ICT Using Three Independent Samples

While a few respondents in a population may opt to respond using an RR device, others may not find the variable sensitive enough and may express willingness for providing a DR. On the other hand, other respondents may choose to respond using Item Count Technique. To accommodate all these three types of responses, an ORR device has been proposed in the section, wherein, the respondents are free to choose any of the three mediums of answering *viz.*, DR, RR or ICT. If a respondent i in the first sample s_1 opts DR, then, he/ she has to just answer the true value y_i. For providing an RR, the respondent must multiply y_i with a'_{11i}, randomly chosen from $(d_1, d_2, d_3, \dots, d_T)$, with mean $\mu_a = \frac{1}{T}\sum_{t=1}^{T} d_t = 1$ and add the resulting number with twice a number, b'_{11i} randomly chosen from $(f_1, f_2, f_3, \dots, f_M)$, with mean $\mu_b = \frac{1}{M}\sum_{m=1}^{M} f_m$. For simplicity, Device- I *vide* Chaudhuri (2011) and Chaudhuri and Christofides (2013), is used for getting RR responses. The ICT questionnaire comprises of G innocuous quantitative questions with the $(G+1)^{th}$ item on value of y_i. Using ICT, one has to answer the sum of all questions, say t'_{1i}, where,

$$t'_{1i} = \sum_{h=1}^{G} u_{ih} + y_i \tag{20}$$

where, $\sum_{h=1}^{G} u_{ih}$ is the sum of the answers to the innocuous questions. Consider the unknown probabilities for the i^{th} respondent opting for DR, RR and ICT are C_{1i}, C_{2i} and $(1 - C_{1i} - C_{2i})$, respectively, with $0 \le C_{1i} \le 1$, $0 \le C_{2i} \le 1$. The probabilities C_{1i} and C_{2i} are assumed to be unknown and different for each individual in the population. Let the i^{th} respondent's answer be,

$$z'_{11i} = \begin{cases} y_i \; with\; probability\; C_{1i} \\ (a'_{11i}y_i + 2b'_{11i}) \; with\; probability\; C_{2i}\,, 0 \le C_{1i} \le 1, 0 \le C_{2i} \le 1 \\ t'_{1i} \; with\; probability\; (1 - C_{1i} - C_{2i}) \end{cases} \tag{21}$$

In a similar way, the respondent i is requested to provide another response, z'_{12i}, independent from z'_{11i}. Denoting a'_{12i} as a number chosen from $(d_1, d_2, d_3, \dots, d_T)$ and b'_{12i} from $(f_1, f_2, f_3, \dots, f_M)$ by the respondent, independent of the selection of a'_{11i} and b'_{11i}, respectively, then,

$$z'_{12i} = \begin{cases} y_i \ with\ probability\ C_{1i} \\ (a'_{12i} y_i + 2b'_{12i})\ with\ probability\ C_{2i}, 0 \le C_{1i} \le 1, 0 \le C_{2i} \le 1 \\ t'_{1i}\ with\ probability\ (1 - C_{1i} - C_{2i}) \end{cases} \tag{22}$$

Then, proceeding similar to (6), let,

$$r'_{1i} = \frac{z'_{11i} + z'_{12i}}{2}, \quad v'_{1i} = \frac{(z'_{11i} - z'_{12i})^2}{4} \tag{23}$$

Then,

$$E_R(r'_{1i}) = E_R(z'_{11i}) = E_R(z'_{12i}) \tag{24}$$

and,

$$E_R(v'_{1i}) = V_R(r'_{1i}) \tag{25}$$

A respondent k in the second sample s_2, can choose to answer either directly or answer as per an RR device or answer according to an ICT questionnaire, without revealing the choice. However, the RR device and the ICT questionnaire are designed differently. For providing an RR, the respondent is requested to choose a random number, say b'_{21k}, from $(f_1, f_2, f_3, \dots, f_M)$. The third option *i.e.*, the questionnaire for ICT bears the same G innocuous quantitative questions as in the questionnaire used for the first sample, with another set of H innocuous quantitative questions. Let the sum of the answers in the questionnaire be t'_{2k}, where,

$$t'_{2k} = \sum_{h=1}^{G} u_{kh} + \sum_{o=1}^{H} w_{ko} \tag{26}$$

where, $\sum_{h=1}^{G} u_{kh}$ and $\sum_{o=1}^{H} w_{ko}$ are the sum of answers to the G and H innocuous statements, respectively, in the questionnaire. Let the k^{th} respondent's answer as z'_{21k},

$$z'_{21k} = \begin{cases} y_k\ with\ probability\ C_{1k} \\ b'_{21k}\ with\ probability\ C_{2k} \\ t'_{2k}\ with\ probability\ (1 - C_{1k} - C_{2k}) \end{cases}, 0 \le C_{1k} \le 1, 0 \le C_{2k} \le 1 \tag{27}$$

Another response, z'_{22k} is collected from the respondent k, independent from z'_{21k}, following the same procedure. Let b'_{22k} denote the number randomly chosen from $(f_1, f_2, f_3, \dots, f_M)$ and independent of b'_{21k}. Then,

$$z'_{22k} = \begin{cases} y_k\ with\ probability\ C_{1k} \\ b'_{22k}\ with\ probability\ C_{2k} \\ t'_{2k}\ with\ probability\ (1 - C_{1k} - C_{2k}) \end{cases}, 0 \le C_{1k} \le 1, 0 \le C_{2k} \le 1 \tag{28}$$

Then, using (6), consider,

$$r'_{2k} = \frac{z'_{21k} + z'_{22k}}{2}, \quad v'_{2k} = \frac{(z'_{21k} - z'_{22k})^2}{4} \tag{29}$$

Then,

$$E_R(r'_{2k}) = E_R(z'_{21k}) = E_R(z'_{22k}) \tag{30}$$

and,
$$E_R(\, v'_{2k}) = V_R(\, r'_{2k}) \tag{31}$$

A third sample s_3 is selected (independent of the selection of s_1 and s_2) from U with a pre-assigned probability $p(s_3)$ according to an unequal probability sampling design P admitting positive first order and second order inclusion-probabilities $\pi_d = \sum_{s_3 \ni d} P(s_3)$, $\pi_{dq} = \sum_{s_3 \ni d,q} P(s_3), d \neq q, (d, q = 1,2,3, \dots, N)$. Respondents have three types of options *viz.*, DR, RR and ICT for providing their response by choosing any one and answering accordingly. Unlike the first and second samples, usage of the RR device mandates a respondent to select a number, say b'_{31d} randomly from $(f_1, f_2, f_3, \dots, f_M)$, multiply it with "- 1" and then report the resulting answer. Here, the ICT questionnaire contains H innocuous quantitative questions which are same as those in the second questionnaire. Consider the sum of answers in the questionnaire is t'_{3d} for the respondent d, where

$$t'_{3d} = \sum_{o=1}^{H} w_{do} \tag{32}$$

where, $\sum_{o=1}^{H} w_{do}$ is the sum of answers of all the H questions. Let the d^{th} respondent's answer be z'_{31d},

$$z'_{31d} = \begin{cases} y_d \text{ with probability } C_{1d} \\ -b'_{31d} \text{ with probability } C_{2d} \\ t'_{3d} \text{ with probability } (1 - C_{1d} - C_{2d}) \end{cases}, 0 \le C_{1d} \le 1, 0 \le C_{2d} \le 1 \tag{33}$$

This individual is asked to provide another such response, say z'_{32d}, independent from z'_{31d}. With b'_{32d} as the number chosen from $(f_1, f_2, f_3, \dots, f_M)$, independent of the selection of b'_{31d},

$$z'_{32d} = \begin{cases} y_d \text{ with probability } C_{1d} \\ -b'_{32d} \text{ with probability } C_{2d} \\ t'_{3d} \text{ with probability } (1 - C_{1d} - C_{2d}) \end{cases}, 0 \le C_{1d} \le 1, 0 \le C_{2d} \le 1 \tag{34}$$

Then, proceeding similar to (6), consider,

$$r'_{3d} = \frac{z'_{31d} + z'_{32d}}{2}, \quad v'_{3d} = \frac{(z'_{31d} - z'_{32d})^2}{4} \tag{35}$$

Then,
$$E_R(r'_{3d}) = E_R(z'_{31d}) = E_R(z'_{32d}) \tag{36}$$

and,
$$E_R(v'_{3d}) = V_R(r'_{3d}) \tag{37}$$

Consider the estimator e', where,

$$e' = \frac{1}{N} \sum_{i \in s_1} \frac{r'_{1i}}{\pi_i} - \frac{1}{N} \sum_{k \in s_2} \frac{r'_{2k}}{\pi_k} + \frac{1}{N} \sum_{d \in s_3} \frac{r'_{3d}}{\pi_d} \tag{38}$$

Then, it can be easily shown that,

$$E(e') = E_R E_P(e') = E_P E_R(e') = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{Y} \tag{39}$$

Hence, e' is an unbiased estimator of $\bar{Y}$. Then, following Chaudhuri and Pal (2002), variance of e' is,

$$V(e') = E_R\left[\frac{1}{N^2}\left\{\sum_{i}^{N}\sum_{<j}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{r'_{1i}}{\pi_i} - \frac{r'_{1j}}{\pi_j}\right)^2 + \sum_{i=1}^{N}\frac{\beta_i}{\pi_i}{r'_{1i}}^2\right\}\right] + \frac{1}{N^2}\sum_{i=1}^{N} V_R(r'_{1i})$$
$$+E_R\left[\frac{1}{N^2}\left\{\sum_{k}^{N}\sum_{<l}^{N}(\pi_k\pi_l - \pi_{kl})\left(\frac{r'_{2k}}{\pi_k} - \frac{r'_{2l}}{\pi_l}\right)^2 + \sum_{k=1}^{N}\frac{\beta_k}{\pi_k}{r'_{2k}}^2\right\}\right] + \frac{1}{N^2}\sum_{k=1}^{N} V_R(r'_{2k}) \tag{40}$$
$$+E_R\left[\frac{1}{N^2}\left\{\sum_{d}^{N}\sum_{<q}^{N}(\pi_d\pi_q - \pi_{dq})\left(\frac{r'_{3d}}{\pi_d} - \frac{r'_{3q}}{\pi_q}\right)^2 + \sum_{d=1}^{N}\frac{\beta_d}{\pi_d}{r'_{3d}}^2\right\}\right] + \frac{1}{N^2}\sum_{d=1}^{N} V_R(r'_{3d})$$

where,

$$\beta_i = 1 + \frac{1}{\pi_i}\sum_{j\neq i}^{N}\pi_{ij} - \sum_{i=1}^{N}\pi_i\,, \qquad \beta_k = 1 + \frac{1}{\pi_k}\sum_{k\neq l}^{N}\pi_{kl} - \sum_{k=1}^{N}\pi_k$$

and,

$$\beta_d = 1 + \frac{1}{\pi_d}\sum_{d\neq q}^{N}\pi_{dq} - \sum_{d=1}^{N}\pi_d$$

If every sample s_1, s_2 and s_3 contains a common number of distinct units in it, then, $\beta_i = 0\ \forall$ i, $\beta_k = 0\ \forall$ k and $\beta_d = 0\ \forall\ d$ in $V(e)$ above. Then, an unbiased estimator of $V(e')$ is given by,

$$v(e') = \frac{1}{N^2}\left\{\sum_{i<}\sum_{j\in s_1}\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{r'_{1i}}{\pi_i} - \frac{r'_{1j}}{\pi_j}\right)^2 + \sum_{i\in s_1}\frac{\beta_i}{{\pi_i}^2}{r'_{1i}}^2\right\} + \frac{1}{N^2}\sum_{i\in s_1}\frac{v'_{1i}}{\pi_i}$$

$$+\frac{1}{N^2}\left\{\sum_{k<}\sum_{l\in s_2}\left(\frac{\pi_k\pi_l - \pi_{kl}}{\pi_{kl}}\right)\left(\frac{r'_{2k}}{\pi_k} - \frac{r'_{2l}}{\pi_l}\right)^2 + \sum_{k\in s_2}\frac{\beta_k}{{\pi_k}^2}{r'_{2k}}^2\right\} + \frac{1}{N^2}\sum_{k\in s_2}\frac{v'_{2k}}{\pi_k} \tag{41}$$

$$+\frac{1}{N^2}\left\{\sum_{d<}\sum_{q\in s_3}\left(\frac{\pi_d\pi_q - \pi_{dq}}{\pi_{dq}}\right)\left(\frac{r'_{3d}}{\pi_d} - \frac{r'_{3q}}{\pi_q}\right)^2 + \sum_{d\in s_3}\frac{\beta_d}{{\pi_d}^2}{r'_{3d}}^2\right\} + \frac{1}{N^2}\sum_{d\in s_3}\frac{v'_{3d}}{\pi_d}$$

with $\beta_i = 0\ \forall$ i, $\beta_k = 0\ \forall$ k and $\beta_d = 0\ \forall$ d in $v(e')$ when applicable. Hence, $v(e')$ is an unbiased estimator of $V(e')$, such that, $E\{v(e')\} = E_P E_R\big(v(e')\big) = E_R E_P\big(v(e')\big) = V(e')$. A $100(1-\alpha)\%$ Confidence Interval for $\bar{Y}$ is, $[L', U']$, where,

$$L' = e' - \left(\tau_{\alpha/2}\sqrt{v(e')}\right), \qquad U' = e' + \left(\tau_{\alpha/2}\sqrt{v(e')}\right) \tag{42}$$

4. Numerical illustration

To examine the performance of the proposed ORRT devices in Sections 2 and 3, a simulated population of $N = 117$ individuals has been considered, wherein, the variable y indicates amount of tax evaded in the previous financial year and z is the number of family members of the respondent. To estimate the mean value of y, *i.e.*, $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, two independent samples each of size 13 individuals are drawn by employing the Hartley and Rao (1962) sampling method, using z (correlation with y is 0.686) as the size measure for sampling the units. For the using the third devices, three independent samples, each of size 13 are drawn following the same sampling scheme.

The specifications of the RR device followed are stated below:

(i) ORR with two options (Section 2) – RR and ICT using two independent samples

An individual in the first sample, opting for RR has to choose a random number from $(0.498, 0.518, -0.004, 1.501, 1.938, 0.968, 1.414, 1.416, 0.425, 2.507)$ which are pre-fixed by the investigator such that the mean of the numbers is 1. Then the individual has to multiply the chosen number with his/her y-value and add the resulting number with a value randomly chosen from $(-0.036, 1.930, 3.463, 2.253, 3.660, 1.717, -0.047, 1.728, 3.031, 2.928, 1.681)$ which are pre-fixed by the investigator. An individual in the second sample, opting for RR has to choose a random number from $(-0.036, 1.930, 3.463, 2.253, 3.660, 1.717, -0.047, 1.728, 3.031, 2.928, 1.681)$.

(ii) ORR with three options (Section 3) – DR, RR and ICT using three independent samples

An individual in the first sample who chooses to provide a RR is requested to choose a random number from $(2.103, 0.126, 0.307, 1.807, -1.601, 1.236)$ which are pre-fixed by the investigator such that the mean of the numbers is 1. Then the individual has to multiply the chosen number with the true y -value and add the number with a random number from $(6.624, 4.280, 8.512, 2.421, 5.955, 3.237, -2.180, 10.839)$ which are pre-fixed by the investigator. An individual in the second sample choosing RR device, is required to choose a number from $(6.624, 4.280, 8.512, 2.421, 5.955, 3.237, -2.180, 10.839)$. In the third sample, respondents opting to answer using an RR device, are instructed to choose a number from $(6.624, 4.280, 8.512, 2.421, 5.955, 3.237, -2.180, 10.839)$, multiply with “- 1” and then report the resulting answer.

Following the specifications of the ICT, two sets of $G = 5$ and $H = 4$ innocuous statements denoted by B_1, B_2, B_3, B_4, B_5 and E_1, E_2, E_3, E_4, are:

Set-1

B_1 : Number of times I brush my teeth daily.

B_2 : Number of coaching classes attended by my children in a week.

B_3 : Number of chairs in my home.

B_4 : Number of days I exercise in a week.

B_5 : Number of visits to the neighbor market in a week.

Set-2

E_1 : Number of visits to the doctor during the last week.

E_2 : Number of rooms in my house.

E_3 : Number of leaves taken from office during the past six months.

E_4 : Volume (in litres) of milk being purchased daily.

The ICT questionnaires are described below:

(i) ORR with two options (Section 2) – RR and ICT using two independent samples

For both the samples, G innocuous item statements in the questionnaire are given in Set-1. The $(G+1)^{th}$ item in the ICT questionnaire used for the first sample is "Number of occasions at which tax payment was evaded".

(ii) ORR with three options (Section 3) – DR, RR and ICT using three independent samples

G innocuous statements in Set-1 are present in the questionnaire for the first sample. The $(G+1)^{th}$ item in the ICT questionnaire used for the first sample is "Number of occasions at which tax payment was evaded". Questionnaire for the second sample contains $(G+H)$ innocuous statements given in Set-1 and Set-2. H innocuous statements used for the third sample are given in Set-2.

Various scenarios on respondents' choices for DR, RR and ICT for the two devices are identified. For each of these scenarios, e, $v(e)$, L and U for ORR with two options and e', $v(e')$, L' and U' for ORR with three options are computed. The estimates are derived each time for $D = 1000$ re-samples drawn from the population and then the following are calculated:

Average Estimate (AE): $\frac{1}{D}\sum_{d=1}^{1000} e_d$ *and* $\frac{1}{D}\sum_{d=1}^{1000} e'_d$,

Average Variance Estimate (AVE): $\frac{1}{D}\sum_{d=1}^{1000} v(e_d)$ and $\frac{1}{D}\sum_{d=1}^{1000} v(e'_d)$,

Average Coefficient of Variation (ACV): $\frac{1}{D}\sum_{d=1}^{1000} \frac{\sqrt{v(e_d)}}{e_d} 100\%$ *and* $\frac{1}{D}\sum_{d=1}^{1000} \frac{\sqrt{v(e'_d)}}{e'_d} 100\%$

Average Relative Bias (ARB): $\left|\frac{\frac{1}{D}\sum_{d=1}^{1000} e_d - \bar{Y}}{\bar{Y}}\right|$ *and* $\left|\frac{\frac{1}{D}\sum_{d=1}^{1000} e'_d - \bar{Y}}{\bar{Y}}\right|$

Actual Coverage Percentage (ACP) *i.e.*, percentage of cases out of 1,000 re-samples, in which (L, U) and (L', U') covers $\bar{Y}$ and Average Length (AL) of the 1,000 replicates of Confidence Intervals for $\bar{Y}$ for both the devices are also computed. Using Chaudhuri (2018), an ACV less than 10% denotes that the estimate is excellent, $10\% < ACV \leq 20\%$ indicates that the estimate is ok, $20\% < ACV \leq 30\%$ signifies that the estimate is poor, but allowable but the estimate should be discarded if $ACV > 30\%$. An efficient device would also result in lower values of ARB, ACP nearer to 95% and smaller values of AL. The values of AE, AVE, ACV, ARB, ACP and AL obtained from the two devices are displayed in Tables 1 and 2.

Table 1: Performance of ORR Device with options for RR and ICT

Sample Proportion of Individuals Choosing a Response Option		AE (in Rs. lakh) $\bar{Y} = Rs. 1.987$ lakh	AVE	ACV (%)	ARB	ACP (%)	AL (in Rs. lakh)
RR	ICT						
0.909	0.091	2.290	3.429	33.222	0.642	99	6.655
0.636	0.364	2.435	13.067	0.443	1.241	100	12.534
0.909	0.091	2.029	8.116	16.905	0.110	100	7.898
0.556	0.444	2.115	11.706	6.123	0.341	100	12.046
0.889	0.111	2.379	3.848	36.100	0.140	99	6.798
0.875	0.125	2.097	6.076	6.825	0.021	96	7.295
0.500	0.500	2.222	5.406	11.771	0.243	99	7.495
0.667	0.333	1.641	10.097	2.357	0.843	100	8.313
0.333	0.667	1.548	11.529	23.077	0.227	100	10.084

Table 2: Performance of Generalized ORR Device with options for DR, RR and ICT

Sample Proportion of Individuals Choosing a Response Option			AE (in Rs. lakh) $\bar{Y} = Rs. 1.987$ lakh	AVE	ACV (%)	ARB	ACP (%)	AL (in Rs. lakh)
DR	RR	ICT						
0.077	0.615	0.308	1.655	39.765	26.602	1.653	99	21.276
0.154	0.692	0.154	1.667	23.721	27.539	1.181	100	17.421
0.308	0.077	0.615	1.223	85.670	12.160	0.657	100	30.845
0.692	0.154	0.154	2.166	19.790	9.707	0.159	100	15.087
0.308	0.154	0.538	1.614	65.948	1.541	0.020	100	27.501
0.385	0.231	0.385	1.988	42.532	49.269	0.720	100	22.911
0.538	0.154	0.308	1.874	40.188	41.456	2.099	100	20.861
0.615	0.308	0.077	1.289	14.653	5.095	2.409	100	13.099

From Table 1, it is observed that the ORR device with two options (RR and ICT) performs well, as the AEs are close to the population mean $\bar{Y} = Rs. 1.987$ lakh, the ACV values are mostly on the lower side and the ACP is generally above 95%. For the results displayed in Table 2 regarding the ORR device with three options (DR, RR and ICT), the observations are similar to those for those in Table 1. Hence, it can be concluded that both the proposed ORR devices perform well in estimating the population mean of a sensitive variable.

5. Conclusion

To estimate the finite population mean of a sensitive quantitative variable, a practical issue with Randomized Response devices is that some of the respondents may not find the variable sensitive and may insist on answering the true value of the variable. This paper at first proposes an Optional Randomized Response device with response options for Randomized Response and Item Count Technique, to estimate a finite population mean of a quantitative sensitive variable. Next, anticipating that while a few respondents may not hesitate to provide a Direct Response, rest of the individuals may like to opt for a Randomized Response or answer using an Item Count questionnaire. Hence, this paper proposes another Optional Randomized Response device with all the three response options *viz.*, Direct Response, Randomized Response and Item Count Technique. This device has been proposed with the motive to accommodate various response choices of individuals and thus to avoid non-responses from potential respondents. Further, instead of restricting the selection of samples by simple random sampling, this device is designed in such a way that the samples can be drawn from the population by using any general sampling scheme. Due to these reasons, the device is unique in itself and not comparable with the Optional Randomized Response devices existing in the literature. Based on a simulation exercise both the devices perform well in terms of Average Coefficient of Variance, Average Relative Bias, Average Coverage Percentage and Average Length of estimated confidence intervals, thus providing evidence on the possible usefulness of the proposed devices.

Acknowledgements

The authors are thankful to the reviewer for providing valuable comments which have helped in improving the manuscript.

References

[1] Adhikari, A. K., Chaudhuri, A. and Vijayan, K. (1984): Optimum Sampling Strategies for Randomized Response Trials. International Statistical Review 52(2): 115–125.

[2] Arnab, R. (2004): Optional randomized response techniques for complex designs. Biometrical Journal 46: 114-124.

[3] Arnab, R. (2018): Optional randomized response techniques for quantitative characteristics. Communications in Statistics – Theory and Methods 48(16): 4154-4170.

[4] Arnab, R. and Rueda, M. (2016): Optional Randomized Response: A Critical Review. Handbook of Statistics on Data Gathering, Analysis and Protection of Privacy through Randomized Response Techniques: Qualitative and Quantitative Human Traits 34: 387-403.

[5] Chaudhuri, A. (1987): Randomized Response Surveys of Finite Populations: A Unified Approach with Quantitative Data. Journal of Statistical Planning and Inference 15: 157–165.

[6] Chaudhuri, A. (1992): Randomized Response: Estimating Mean Square Errors of Linear Estimators and Finding Optimal Unbiased Strategies. Metrika 39: 341–357.

[7] Chaudhuri, A. (2011): Randomized response and indirect questioning techniques in surveys. Boca Raton: Chapman & Hall, CRC Press, Taylor & Francis Group.

[8] Chaudhuri, A. (2018): Survey Sampling. Chapman & Hall: CRC Press, Taylor and Francis Group.

[9] Chaudhuri, A. and Christofides, T. C. (2007): Item count technique in estimating the proportion of people with a sensitive feature. Journal of Statistical Planning and Inference 137: 589-593.

[10] Chaudhuri, A. and Christofides, T. C. (2013): Indirect Questioning in Sample Surveys. Springer.

[11] Chaudhuri, A. and Dihidar, K. (2009): Estimating means of stigmatizing qualitative and quantitative variables from discretionary responses randomized or direct. Sankhya B71: 123-136.

[12] Chaudhuri, A. and Mukherjee, R. (1985): Optionally randomized response techniques. Calcutta Statistical Association Bulletin 34: 225-229.

[13] Chaudhuri, A. and Pal, S. (2002): On certain alternative mean square error estimators in complex survey sampling. Journal of Statistical Planning and Inference 104: 363-375.

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

[15] Eriksson, S. A. (1973): A New Model for Randomized Response. International Statistical Review 41(1): 101–113.

[16] Fox, J. A. and Tracy, P. E. (1984): Measuring Associations with Randomized Response. Social Science Research 13(2): 188–197.

[17] Greenberg, B. G., Kuebler, R. R., Abernathy, J. R. and Horvitz, D. G. (1971): Application of the Randomized Response Technique in Obtaining Quantitative Data. Journal of the American Statistical Association 66(334): 243–250.

[18] Gupta, S., Shabbir, J. and Sehra, S. (2010): Mean and sensitivity estimation in optional randomized response models. Journal of Statistical Planning and Inference 140: 2870–2874.

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

[20] Hartley, H.O. and Rao, J. N. K. (1962): Sampling with unequal probabilities and without replacement. Annals of Mathematical Statistics 33(2): 350-374.

[21] Miller, J.D. (1984): A new survey technique for studying deviant behaviour. Ph.D. Thesis, The George Washington University.

[22] Miller, J. D., Cisin, I. H. and Harrel, A. V. (1986): A new technique for surveying deviant behavior: item count estimates of marijuana, cocaine and heroin. Paper presented at the annual meeting of the American Association for Public Opinion Research, St. Petersburg, Florida.

[23] Pal, S. (2007): Estimating the proportion of people bearing a sensitive issue with an option to item count lists and randomized response. Statistics in Transition 8(2): 301-310.

[24] Pal, S. (2008): Unbiasedly estimating the total of a stigmatizing variable from a complex survey on permitting options for direct or randomized responses. Statistical Papers 49(2): 157-164.

[25] Pal, S., Chaudhuri, A. and Patra, D. (2020): How privacy may be protected in optional randomized response surveys. Statistics in Transition 21(2): 61-87.

[26] Raghavarao, D. and Federer, W.F. (1979): Block total response as an alternative to the randomized response method in surveys. Journal of the Royal Statistical Society Series B, 41: 40-45.

[27] Sengupta, S. and Kundu, D. (1989): Estimation of Finite Population Mean in Randomized Response Surveys. Journal of Statistical Planning and Inference 23(1): 117–125.

[28] Shaw, P. (2015): Estimating a finite population mean of a sensitive quantitative variable from a single probability sample by the Item Count Technique. Model Assisted Statistics and Applications 10: 411-419.

[29] Shaw, P. and Pal, S. (2021): Estimating Sensitive Population Proportion Permitting Options for Various Respondents' Choices. Statistics and Applications 19(2): 161-179.

[30] Trappmann, M., Krumpal, I., Kirchner, A. and Jann, B. (2014): Item sum: a new technique for asking quantitative sensitive questions. Journal of Survey Statistics and Methodology 2: 58-77.

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