STATISTICS IN TRANSITION new series, September 2023
Vol. 24, No. 4, pp. 53–70, https://doi.org/10.59170/stattrans-2023-051
Received – 16.11.2021; accepted – 19.10.2022

Respondent-specific randomized response technique to estimate sensitive proportion

Dipika Patra[1], **Sanghamitra Pal**[2], **Arijit Chaudhuri**[3]

Abstract

In estimating the proportion of people bearing a stigmatizing characteristic in a community of people, randomized response techniques are plentifully available in the literature. They are implemented essentially using boxes of similar cards of two distinguishable types. In this paper, we propose a more general procedure using five different types of cards. A respondent-specific randomized response technique is also proposed, in which respondents are allowed to build up the boxes according to their own choices. An immediate objective for this change is to enhance, sense of protection of privacy of the respondents. But as by-products, higher efficiency in terms of actual coverage percentages of confidence intervals and related features are demonstrated by a simulation study, and superior jeopardy levels against divulgence of personal secrecy are also reported to be achievable.

AMS subject classification: 62D05

Key words: protection of privacy, randomized response, sensitive issues, varying probability sampling.

1. Introduction

Paying heed to Chaudhuri's (2011) text, we consider varying probability sampling designs in surveying finite survey population with the purpose of estimating the proportion of people bearing a sensitive feature like tax evasion, bearing criminal antecedents, etc. in a community of persons. Randomized Response (RR) Techniques (RRTs) with standard procedures given by Warner (1965), Simmons (his URL model), Boruch (his Forced Response model) (vide Chaudhuri (2011) for each) and others are well-known and documented. They essentially employ boxes filled with several

[1] Department of Statistics, Seth Anandram Jaipuria College, Kolkata, India.
E-mail: dipika.patra1988@gmail.com. ORCID: https://orcid.org/0000-0003-4318-1123.
[2] Corresponding author. Department of Statistics, West Bengal State University, India.
E-mail: mitrapal2013@gmail.com. ORCID: https://orcid.org/0000-0002-5752-8282.
[3] Applied Statistics Unit, Indian Statistical Institute, Kolkata, India. E-mail: arijitchaudhuri1@rediffmail.com.
ORCID: https://orcid.org/0000-0002-4305-7686.

identically designed cards with two distinct types of visible marks. Standard procedures of unbiased estimation of the proportion of people bearing the sensitive characteristic, say, A, along with variance formulae and unbiased estimators thereof are available in the above location cited. In recent surveys also, RR technique is quite popular. Treating illegal waste disposal as a sensitive attribute, Chong et al. (2019) analyzed the social problem of waste disposal with RR technique. Arnab and Mothupi (2015) assessed the sexual habits of the University Students, using Warner's (1965) and Greenberg et al.'s (1969) RR techniques. Barabesi et al. (2013) employed RR setups for the estimation of the size of hidden gang and the distribution function of a sensitive variable for the members of the group. Van der Heijden et al. (2000) applied the Forced Response technique and Kuk's (1990) RR technique to obtain reliable data on welfare and unemployment benefits fraud which is highly relevant to policy decisions.Together with various applications of RRT, statistical tools were also developed to analyze RR data. For instance, Hout et al. (2007) discussed the univariate and multivariate logistic regression models to measure sensitive feature. They presented univariate model as a generalized linear model and introduced multivariate model to deal with several RR response variables. Also, Fox et al. (2018) considered a generalized linear model and generalized linear mixed model for RR design. The literature to be cited below is rich giving procedures to provide methods and measures of levels of protection verifiable for the respondents' disclosures of privacy.

In the existing literature of RRT, the interviewer constructs the RR device(s) and the respondents are requested to participate in the RR survey. In practice, respondents hesitate to participate in RR survey. Anticipating more participation in such survey, a new RR survey theory has been proposed, in which respondents are allowed to construct the RR devices. This proposed RR technique is termed as *respondent-specific randomized response technique*.

The paper is organized as follows. Section 2 provides certain basics for RRT in the context of qualitative sensitive features. A brief description of the protection of privacy measures is included there. Section 3 is constructed to propose two general RR techniques covering varying probability sampling design. Section 3.1 describes Model 1 in which five distinct types of cards are used in RR device. Section 3.2 proposes a novel RR device in which respondents are asked to build up the RR boxes according to their own choices. Section 4 is devoted to the measure the respondents' privacy protection. Privacy is protected only for a RR-specific parametric combination and such a feature will be seen in this section. The effectiveness and competitiveness of the proposed RRTs are narrated through numerical findings, in Section 5. This article is ended with some concluding remarks in Section 6.

2. The Early Works

Taking a cue from the pioneering work of Warner (1965), Greenberg et al. (1969) recommended unrelated question model with two questions of which one is about the sensitive characteristic A and the other question is unrelated to the sensitive characteristic. The idea of this RR device is originated by Walt R. Simmons. The reason behind this extension is that like A, it is a complement, i.e. A^c may be a sensitive characteristic. In that case, the respondents may hesitate to give out their true nature. Chaudhuri (2011) developed the RR devices and the estimation procedures for general sampling design. The extensions of the work are narrated in Chaudhuri (2011) (chapter 3), Chaudhuri et al. (2016). Boruch's (1972) Forced RRT considers the RR device with three distinct types of cards. Instead of the unrelated question, he suggested to include the cards marked as "Yes" and "No". Taking a cue from them, a new RR technique has been suggested in this paper with five different options in the RR device. In another proposed RR technique, respondents are allowed to build their RR devices choosing different cards according to their own choices. Then, the respondents will be comfortable to participate in RR survey.

Several authors including Lanke (1975,1976), Leysieffer and Warner (1976), Anderson (1975 a,b,c), Diana and Perii (2013) have drawn the attention of many survey practitioners to measure the degree of protection of the responses. However, their measures are confined to Simple Random Sampling (SRS) with replacement. Chaudhuri, Christofides and Saha (2009) covered the protection of privacy measure for RRTs using general sampling design. With the approach of Chaudhuri et al. (2009) and Pal et al. (2020) the protection of privacy measure has been derived here for the proposed generalized RR techniques.

3. Proposed RR Techniques Using Five-types of Cards

A potentially useful generalized RRT is proposed in sub-section 3.1 as Model 1. Additionally, a respondent-specific randomized response technique is also introduced in the sub-section 3.2 as Model 2.

3.1. Model 1: Generalized RR technique

An ameliorated RR technique is proposed here employing two boxes filled with several identically designed cards with 5 distinct types of visible marks as, " I possess A", "I possess A^c", "I possess innocuous character B", "Yes" and "No" having proportions $p_k, (1-p_k)w_2, (1-p_k)w_3, (1-p_k)w_4$ and $(1-p_k)(1-w_2-w_3-w_4)$ respectively in the $k^{th}(k=1,2)$ box $(p_1 \neq p_2, w_2+w_3+w_4 < 1,\ 0 < p_1,\ p_2,\ w_2,\ w_3,\ w_4 < 1)$.

Let $U = (1,2 \ldots N)$ be a finite population on which the variables y and x are defined. The variables y and x are introduced relating to the sensitive attribute A and the innocuous characteristic B respectively.

Thus, for i^{th} $(i \in U)$ person,

$$y_i = \begin{cases} 1 & ,if\ i^{th}\ person\ bears\ A \\ 0 & ,if\ i^{th} person\ bears\ A^c \end{cases}$$

and

$$x_i = \begin{cases} 1 & ,if\ i^{th}\ person\ bears\ B \\ 0 & ,if\ i^{th} person\ bears\ B^c. \end{cases}$$

The aim is to estimate the population proportion $\theta = \frac{1}{N}\sum_{i=1}^{N} y_i\ ;\ \theta \in [0,1]$.

A sample s of size n is drawn from the population U by any sampling design $P(s)$. A sampled person i $(i \in s, i = 1,2, \ldots, n)$ is requested to draw a card randomly from the 1st box without divulging the card-type. The respondent must give out the truthful response in terms of yes or no according to the card type marked as "I possess A", "I possess A^c" or "I possess innocuous character B". The person is also instructed to report yes or no if the card is marked as "Yes" or "No". Figure 3.1 successfully explains the proposed strategy.

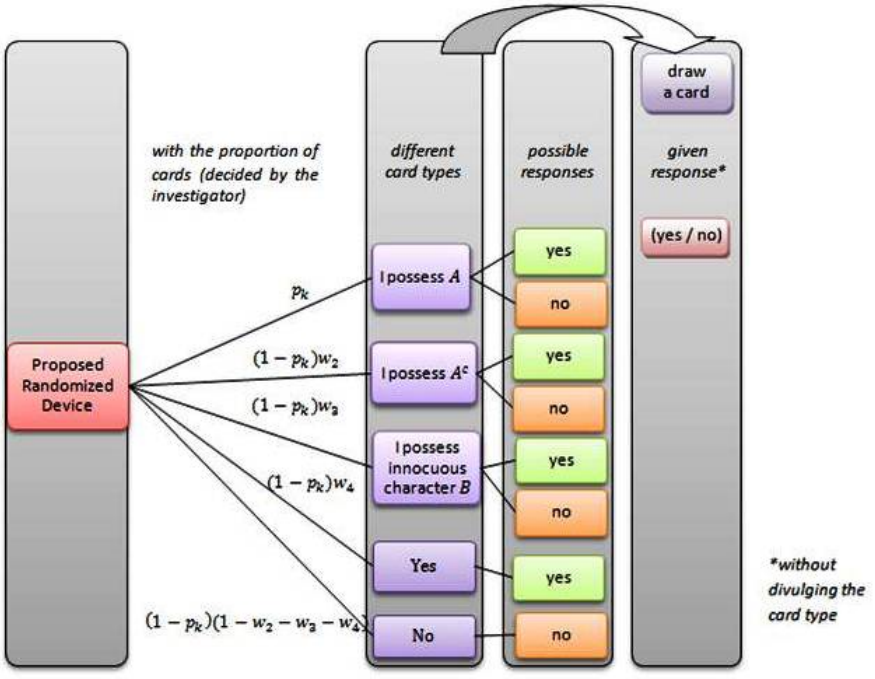

Figure 3.1: Model 1: Generalized RR Technique

Thus, the randomized response from i^{th} $(i \in s)$ person is

$$I_i = \begin{cases} 1 & if\ the\ response\ is\ \ yes \\ 0 & if\ the\ response\ is\ \ no. \end{cases}$$

Therefore,

$$P(I_i = 1) = p_1 y_i + (1 - p_1)\{w_2(1 - y_i) + w_3 x_i + w_4\}$$

and $P(I_i = 0) = p_1(1 - y_i) + (1 - p_1)\{w_2 y_i + w_3(1 - x_i) + (1 - w_2 - w_3 - w_4)\}$.

The person is also requested to report another response described as earlier after drawing a card from the 2nd box, independently.

Therefore, we may denote the 2nd response as,

$$J_i = \begin{cases} 1 & if\ response\ is\ \ yes \\ 0 & if\ response\ is\ \ no \end{cases},\ \ i \in s.$$

Then,

$$P(J_i = 1) = p_2 y_i + (1 - p_2)\{w_2(1 - y_i) + w_3 x_i + w_4\}$$

and $P(J_i = 0) = p_2(1 - y_i) + (1 - p_2)\{w_2 y_i + w_3(1 - x_i) + (1 - w_2 - w_3 - w_4)\}$.

Denoting RR based expectations and variances as E_R and V_R throughout the study, we may write,

$$E_R(I_i) = p_1 y_i + (1 - p_1)\{w_2(1 - y_i) + w_3 x_i + w_4\}$$

and

$$E_R(J_i) = p_2 y_i + (1 - p_2)\{w_2(1 - y_i) + w_3 x_i + w_4\}.$$

Therefore,

$$E_R\big((1 - p_2)I_i - (1 - p_1)J_i\big) = \{p_1(1 - p_2) - p_2(1 - p_1)\}y_i = (p_1 - p_2)y_i$$

$$E_R\left(\frac{(1 - p_2)I_i - (1 - p_1)J_i}{p_1 - p_2}\right) = y_i\,;\quad p_1 \neq p_2$$

leading to

$$r_i = \frac{(1-p_2)I_i - (1-p_1)J_i}{p_1 - p_2},\ p_1 \neq p_2 \tag{1}$$

which is the unbiased estimator for y_i .

An unbiased estimator of the variance $V_R(r_i)$ is given by

$$v_R(r_i) = r_i(r_i - 1) = \frac{(1-p_1)(1-p_2)}{(p_1-p_2)^2}(I_i - J_i)^2, \tag{2}$$

since $y_i^2 = y_i, x_i^2 = x_i, I_i^2 = I_i$ and $J_i^2 = J_i$. The details of the proof are given below.

Considering $v_R^*(r_i) = (1 - p_1)(1 - p_2)(I_i - J_i)^2$ we get,

$$\begin{aligned}
E_R\big(v_R^*(r_i)\big) &= (1 - p_1)(1 - p_2)E_R(I_i - J_i)^2 \\
&= (1 - p_1)(1 - p_2)\big\{E_R\big(I_i^2\big) + E_R\big(J_i^2\big) - 2E_R(I_i)E_R(J_i)\big\} \\
&= (1 - p_2)\{(1 - p_1) - (1 - p_2)\}E_R\big(I_i^2\big) + (1 - p_1)\{(1 - p_2) - (1 - p_1)\}E_R\big(J_i^2\big) \\
&\qquad + E_R\big((1 - p_2)I_i - (1 - p_1)J_i\big)^2 \\
&= (1 - p_2)(p_2 - p_1)E_R\big(I_i^2\big) + (1 - p_1)(p_1 - p_2)E_R\big(J_i^2\big) \\
&\qquad + E_R\big((p_1 - p_2)r_i\big)^2;\ using\ (eq.1) \\
&= -(p_1 - p_2)E_R\big((1 - p_2)I_i - (1 - p_1)J_i\big) + (p_1 - p_2)^2 E_R\big(r_i^2\big);\ since\ I_i^2 = I_i, J_i^2 = J_i \\
&= (p_1 - p_2)^2 E_R\big(r_i^2\big) - (p_1 - p_2)^2 E_R(r_i);\ using\ (eq.\ 1)
\end{aligned}$$

$$= (p_1 - p_2)^2\left(E_R(r_i^2) - y_i\right)$$
$$= \left(p_1 - p_2\right)^2(E_R(r_i^2) - y_i^2);\ since\ y_i = y_i^2$$
$$= (p_1 - p_2)^2(E_R(r_i^2) - (E_R(r_i))^2 = (p_1 - p_2)^2 V_R(r_i).$$

Therefore, $\frac{1}{(p_1-p_2)^2} v_R^*(r_i) = v_R(r_i)$ is the unbiased estimator for $V_R(r_i)$. Employing Horvitz-Thompson (1952) estimator in estimating the population proportion $\theta = \frac{1}{N}\sum_{i\in U} y_i$, the final unbiased estimator can be written as

$$e_{HT} = \frac{1}{N}\sum_{i\in s}\frac{r_i}{\pi_i}. \tag{3}$$

Hence, an unbiased variance estimator is

$$v(e_{HT}) = \frac{1}{N^2}\left[\sum_{i<j\in s}\sum\left(\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\right)\left(\frac{r_i}{\pi_i}-\frac{r_j}{\pi_j}\right)^2 + \sum_{i\in s}\frac{v_R(r_i)}{\pi_i}\right] \tag{4}$$

where $v_R(r_i)$ is derived in (eq.2).

Composition of Such Randomized Device:

Let the sampled persons be approached with two boxes. In the 1st box, let there be m identically designed cards of which m_1, m_2, m_3 and m_4 cards have visible marks as "I possess A", "I possess A^c", "I possess innocuous character B" and "Yes" respectively. The remaining $(m - m_1 - m_2 - m_3 - m_4)$ cards have a mark "No".

Then, the proportion of cards is

"I possess A": "I possess A^c": "I possess B": "Yes": "No"

$$= m_1 : m_2 : m_3 : m_4 : (m - m_1 - m_2 - m_3 - m_4)$$
$$= \frac{m_1}{m} : \frac{m_2}{m} : \frac{m_3}{m} : \frac{m_4}{m} : \frac{(m-m_1-m_2-m_3-m_4)}{m}$$
$$= \frac{m_1}{m} : \left(\frac{m-m_1}{m}\right)\frac{m_2}{m-m_1} : \left(\frac{m-m_1}{m}\right)\frac{m_3}{m-m_1} : \left(\frac{m-m_1}{m}\right)\frac{m_4}{m-m_1} : \left(\frac{m-m_1}{m}\right)\frac{m-m_1-m_2-m_3-m_4}{m-m_1}.$$

Now, taking $\frac{m_1}{m} = p_1$ and $w_j = \frac{m_j}{m-m_1}; j = 2,3,4,$ the above proportion becomes

$p_1 : (1-p_1)w_2 : (1-p_1)w_3 : (1-p_1)w_4 : (1-p_1)(1-w_2-w_3-w_4)$ where $0 < w_j < 1, j = 2,3,4$ and $w_2 + w_3 + w_4 < 1$ are obvious conditions.

The proportion of the above cards in the 2nd box may be done by changing only (adding or removing) a fixed number of "I possess A" marked cards used in the 1st box. Thus, the number of other-types of cards will remain unchanged as in the 1st box.

Then, one can easily see that the proportion of cards in the 2nd box is now changed to

$$p_2 : (1-p_2)\, w_2 : (1-p_2)\, w_3 : (1-p_2)\, w_4 : (1-p_2)(1-w_2-w_3-w_4)$$

Remark:

1. *The proposed generalized RR technique (Model 1) reduces to Warner's (1965) RRT if* $w_2 = 1,\ w_3 = 0,\ w_4 = 0$.
2. *The proposed Model 1 reduces to Greenberg et al.'s (1969) RRT if* $w_2 = 0, w_3 = 1, m_4 = 0.$
3. *The proposed Model 1 reduces to the Forced RRT if* $w_2 = 0,\ w_3 = 0, w_4 < 1.$

3.2. Model 2: Respondent-specific RR Technique

Intending to enhance the sense of responses' privacy, we modify the RR technique recounted in the previous section (Section 3.1) giving freedom to the respondents to construct their RR devices with the same five distinct types of cards as mentioned earlier. In such a situation, this generalized RR technique, termed as ***Respondent-specific RR technique*** is relevant. This is quite possible that respondents possessing the sensitive characteristic A may prefer any other types of cards except "Yes" marked cards.

With the following illustration, the implementation of this procedure can be easily understandable. Also, Figure 3.2 sheds light on the specifications of the RR devices.

Let a sampled person be approached with two empty boxes and a sufficient number of cards marked as earlier. On request, the person has to build up 1st box (Box 1) by inserting total $m\ (fixed)$ number of cards. In the building process of Box 1, the person will put $m_1\ (fixed\ and > 0)$ number of "I possess A" cards. The rest of the $(m - m_1)$ cards are marked other than "I possess A" marked cards. In other words, there is no restriction on the number of "I possess A^c", "I possess innocuous character B", "Yes" and "No" marked cards. The 2nd box (Box 2) should be built up with $(m + a)$ number of cards in total where the number of "I possess A" cards is $(m_1 + a)$ and the remaining cards are present here in the same number as given in the Box 1. The value of "a" should be decided by the interviewer. Then, the respondents are requested to draw a card randomly from each of the boxes and respond accordingly. The reason to fix up $m\ (> 0), m_1 (< m)$ and $a\ (> 0)$ for all respondents by the interviewer is discussed later.

Therefore, the proportions of "I possess A", "I possess A^c", "I possess innocuous character B", "Yes" and "No" marked cards in the 1st box and 2nd box become $p_1 : (1 - p_1)\ \ w_2 : (1 - p_1)\ \ w_3 : (1 - p_1)\ w_4 : (1 - p_1)(1 - w_2 - w_3 - w_4)$ and $p_2 : (1 - p_2)\ \ w_2 : (1 - p_2)\ \ w_3 : (1 - p_2)\ w_4 : (1 - p_2)(1 - w_2 - w_3 - w_4)$ respectively. It is noteworthy that w_2, w_3 and w_4 are unknown to the interviewer and depend on the choice of the sampled person. However, $p_1\ and\ p_2\ (\neq p_1)$ are known to the interviewer due to the fixed values of m, m_1 and a.

Survey practitioners may use computerized RR devices like a virtual picker wheel. With the help of Google form investigators may record only the answers from the

respondents. The Google form should contain links of virtual RR devices and the options "yes" and "no" for each device. Respondents can enter into a particular RR device clicking on the link, mentioned in the form.

For example, the link https://pickerwheel.com/pw?id=LeCbM only allows respondents to click on the spin button of the virtual RR device. Picker wheel will show the choice or the statement while the spinning of the wheel is stopped. The respondent is requested to select option "yes" ("no") in Google form if the selected choice is "yes" ("no"). Otherwise he/she will provide a truthful response in terms of "yes" or "no" according to the statement visible on the computer screen.

Another link, https://pickerwheel.com/pw?id=aawQs is also a virtual RR device which can be edited by the respondents to construct their own RR device in the case of Model 2. But, they should follow the instructions given by investigators, strictly.

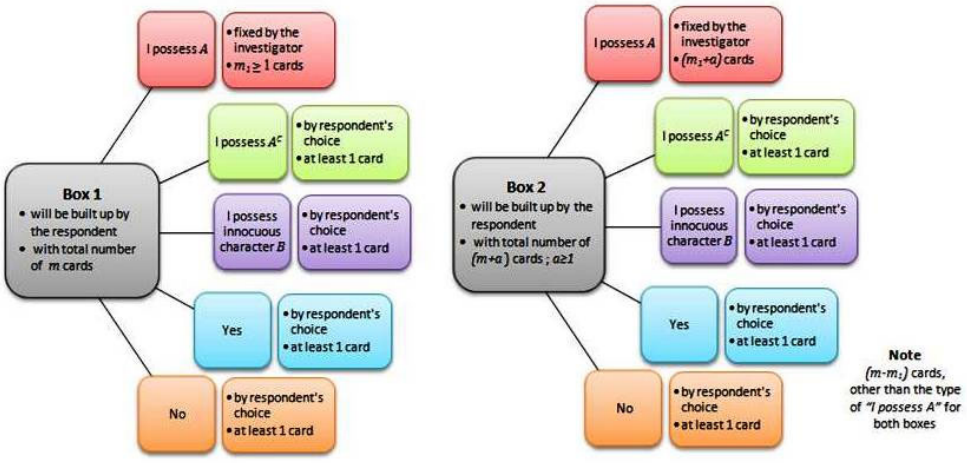

Figure 3.2: Model 2: Respondent-Specific RR Technique

Let I_i' and J_i' be the two independent responses of i^{th} sampled person which are defined as follows:

$$I_i' = \begin{cases} 1 & if\ the\ response\ is\ \ yes \\ 0 & if\ the\ response\ is\ \ no \end{cases}$$

and

$$J_i' = \begin{cases} 1 & if\ the\ response\ is\ \ yes \\ 0 & if\ the\ response\ is\ \ no. \end{cases}$$

Then, taking RR based expectations on I_i' and J_i', we get an unbiased estimator of y_i as

$$r_i' = \frac{(1-p_2)I_i' - (1-p_1)J_i'}{p_1 - p_2};\ \ p_1 \neq p_2\ .$$

Hence, the unbiased estimator of the population proportion θ is given by

$$e'_{HT} = \frac{1}{N}\sum_{i\in s}\frac{r'_i}{\pi_i}.$$

Therefore, the unbiased variance estimator is

$$v(e'_{HT}) = \frac{1}{N^2}\left[\sum_{i<j\in s}\sum\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{r'_i}{\pi_i} - \frac{r'_j}{\pi_j}\right)^2 + \sum_{i\in s}\frac{v_R(r'_i)}{\pi_i}\right]$$

where $v_R(r'_i) = \frac{(1-p_1)(1-p_2)}{(p_1-p_2)^2}(I'_i - J'_i)^2$.

4. Protection of Privacy for the Proposed RRTs

Chaudhuri et al. (2009) investigated the possibility of protecting the privacy of the respondent in RR context using varying probability sampling designs. Pal et al. (2020) recently developed the measure of privacy protection when an opportunity for divulging out the direct response is also given to each respondent along with a specific RR device. The respondents provide either direct response or randomized response, without divulging the type of response so exercised. We refer to Pal et al. (2020) here for a detailed account of the measure of privacy protection with two independent randomized responses.

Denoting the responses of the above respondent specific RR device as (R, R'), the posterior probability and measure of jeopardy of the i^{th} respondent may be written as $L_i(R, R')$ and $J_i(R, R')$ respectively.

Considering the prior probability $L_i\,(0 < L_i < 1)$ for the i^{th} respondent and applying Bayes' theorem, we get

$$Prob\left(A\middle|(R,R')\right) = \frac{L_i P(R|A) P\left(R'\middle|A\right)}{L_i P(R|A) P\left(R'\middle|A\right) + (1-L_i) P(R|A^c) P\left(R'\middle|A^c\right)} = L_i(R,R'). \quad (5)$$

Then, for i^{th} person, the response-specific jeopardy measure

$$J_i(R,R') = \frac{L_i(R,R')/L_i}{1-L_i(R,R')/1-L_i} \quad (6)$$

indicates the risk of divulging the respondent's status due to the responses (R, R').

Consequently, Chaudhuri et al. (2009) and Pal et al. (2020) suggested arithmetic mean $(\bar{J}_i)$ and geometric mean $(\tilde{J}_i)$ respectively as an overall measure of jeopardy.

Here, the possible responses (R, R') are (1,1),(0,0),(1,0) and (0,1).

Then, substituting $(R, R') \equiv (1,1)$ in (eq.5), we get the posterior probability **for the response (1, 1)** as,

$$L_i(1,1) = \frac{L_i P(I_i = 1|y_i = 1) P(J_i = 1|y_i = 1)}{L_i P(I_i = 1|y_i = 1) P(J_i = 1|y_i = 1) + (1 - L_i) P(I_i = 1|y_i = 0) P(J_i = 1|y_i = 0)}$$

$$= \frac{L_i\{p_1 + (1-p_1)(w_3+w_4)\}\{p_2 + (1-p_2)(w_3+w_4)\}}{L_i\{p_1 + (1-p_1)(w_3+w_4)\}\{p_2 + (1-p_2)(w_3+w_4)\} + (1-L_i)(1-p_1)(1-p_2)(w_2+w_3+w_4)^2}. \quad (7.1)$$

and substituting the same in (eq.6), the response specific jeopardy measure **for the response (1, 1)** is

$$J_i(1,1) = \frac{\{p_1+(1-p_1)(w_3+w_4)\}\{p_2+(1-p_2)(w_3+w_4)\}}{(1-p_1)(1-p_2)(w_2+w_3+w_4)^2}. \quad (7.2)$$

Note that $J_i(1,1) \to 1$ if $p_k + (1-p_k)(w_3+w_4) \to (1-p_k)(w_2+w_3+w_4)$; $k = 1,2$.

That imply $p_k \to (1-p_k)w_2$; $k = 1,2$.

i.e. the proportion of "I possess A" cards tending to the proportion of "I possess A^c" cards and $p_1 \to p_2$ are the advisable conditions for protecting the response (1,1) but the condition $p_1 \to p_2$ entails that the variance estimate $v_R(r_i)$ defined in eq. 2 tends to infinite.

Similarly, **for the response (0, 0),**

$$L_i(0,0) = \frac{L_i P(I_i = 0|y_i = 1)P(J_i = 0|y_i = 1)}{L_i P(I_i = 0|y_i = 1)P(J_i = 0|y_i = 1) + (1-L_i)P(I_i = 0|y_i = 0)P(J_i = 0|y_i = 0)}$$

$$= \frac{L_i(1-p_1)(1-p_2)(1-w_4)^2}{L_i(1-p_1)(1-p_2)(1-w_4)^2 + (1-L_i)\{p_1 + (1-p_1)(1-w_2-w_4)\}\{p_2 + (1-p_2)(1-w_2-w_4)\}} \quad (8.1)$$

and

$$J_i(0,0) = \frac{(1-p_1)(1-p_2)(1-w_4)^2}{\{p_1 + (1-p_1)(1-w_2-w_4)\}\{p_2 + (1-p_2)(1-w_2-w_4)\}}. \quad (8.2)$$

The necessary conditions for $J_i(0,0) \to 1$ are $p_k + (1-p_k)(1-w_2-w_4) \to (1-p_k)(1-w_4)$; $\forall\, k = 1,2$. That imply $p_k \to (1-p_k)w_2$; $\forall k = 1,2$.

In other words, the advisable conditions for protecting the response (0,0) are $p_1 \to p_2$ and the proportion of "I possess A" cards tending to the proportion of "I possess A^c" cards. But it entails that $v_R(r_i) \to \infty$.

Now, substituting $(R, R') \equiv (1,0)$ in (eq. 5) and (eq.6), we get

$$L_i(1,0) = \frac{L_i P(I_i = 1|y_i = 1)P(J_i = 0|y_i = 1)}{L_i P(I_i = 1|y_i = 1)P(J_i = 0|y_i = 1) + (1-L_i)P(I_i = 1|y_i = 0)P(J_i = 0|y_i = 0)}$$

$$= \frac{L_i\{p_1 + (1-p_1)(w_3+w_4)\}(1-p_2)(1-w_4)}{L_i\{p_1 + (1-p_1)(w_3+w_4)\}(1-p_2)(1-w_4) + (1-L_i)(1-p_1)(w_2+w_3+w_4)\{p_2 + (1-p_2)(1-w_2-w_4)\}} \quad (9.1)$$

and

$$J_i(1,0) = \frac{\{p_1+(1-p_1)(w_3+w_4)\}(1-p_2)(1-w_4)}{(1-p_1)(w_2+w_3+w_4)\{p_2+(1-p_2)(1-w_2-w_4)\}} \quad (9.2)$$

respectively.

This $J_i(1,0) \to 1$ if $p_1 + (1-p_1)(w_3+w_4) \to (1-p_1)(w_2+w_3+w_4)$ and $p_2 + (1-p_2)(1-w_2-w_4) \to (1-p_2)(1-w_4)$ which imply $p_1 \to (1-p_1)w_2$ and $p_2 \to (1-p_2)w_2$.
i.e. if the proportion of "*I possess A*" cards tends to the proportion of "*I possess* A^c" cards and $p_1 \to p_2$, then $J_i(1,0)$ converges to 1 with $v_R(r_i) \to \infty$.

For the response (0, 1),

$$L_i(1,0) = \frac{L_i P(I_i=0|y_i=1)P(J_i=1|y_i=1)}{L_i P(I_i=0|y_i=1)P(J_i=1|y_i=1) + (1-L_i)P(I_i=0|y_i=0)P(J_i=1|y_i=0)}$$
$$= \frac{L_i\{p_2+(1-p_2)(w_3+w_4)\}(1-p_1)(1-w_4)}{L_i\{p_2+(1-p_2)(w_3+w_4)\}(1-p_1)(1-w_4) + (1-L_i)(1-p_2)(w_2+w_3+w_4)\{p_1+(1-p_1)(1-w_2-w_4)\}} \quad (10.1)$$

and

$$J_i(0,1) = \frac{\{p_2+(1-p_2)(w_3+w_4)\}(1-p_1)(1-w_4)}{(1-p_2)(w_2+w_3+w_4)\{p_1+(1-p_1)(1-w_2-w_4)\}}. \quad (10.2)$$

Now, if $p_k \to (1-p_k)w_2$; $k = 1,2$ then $J_i(0,1)$ converges to 1. But in such a case $v_R(r_i) \to \infty$.

Thus, considering the geometric mean $\tilde{J}_i$ as the overall measure of jeopardy for the proposed RRT, we get

$$\tilde{J}_i = \{J_i(1,1) \times J_i(0,0) \times J_i(1,0) \times J_i(0,1)\}^{1/4}$$
$$= \left\{\frac{p_1+(1-p_1)(w_3+w_4)}{p_1+(1-p_1)(1-w_2-w_4)} \frac{p_2+(1-p_2)(w_3+w_4)}{p_2+(1-p_2)(1-w_2-w_4)} \frac{(1-w_4)^2}{[\![(w]\!]_2+w_3+w_4)^2}\right\}^{1/2}. \quad (11)$$

(eq. 11) converges to 1
if $(1-w_4) \to [\![(w]\!]_2 + w_3 + w_4)$ i.e. $(1-w_2-w_3-w_4) \to w_4$ or

$$\frac{p_k+(1-p_k)(w_3+w_4)}{p_k+(1-p_k)(1-w_2-w_4)} = \frac{(w_2+w_3+w_4)}{1-w_4} \text{ i.e.} p_k \to (1-p_k)w_2 \text{ ; } k = 1,2.$$

In other words, the proposed RR techniques ensure maximum protection if the proportion of "Yes" and "No" cards are the same or the proportion of "I possess A" and "I possess A^c" cards are the same.

Hence, it is advisable to apply this proposed RR techniques with the RR devices having at least one of the following properties:

i) "Yes" and "No" cards are in the same proportion in the devices
ii) "I possess A" and "I possess A^c cards are the same in proportion for both devices.

In Model 1, such restrictions on model parameters may be followed. However, in Model 2 i.e. Respondent-specific RR technique, the restrictions are not guaranteed as RR devices are made by respondents.

In the later section, the measure of jeopardy is calculated numerically for a different combination of p_1 and p_2.

5. Simulation Study

In this section, we investigate the performance of the proposed RRTs using fivetypes of cards through a simulation study. For this, we consider a fictitious data consisting of reckless driving history with weekly expenses of $N = 116$ undergraduate students under 20 years of age. Here, the parameter of interest is the proportion of the students who broke the traffic rules last year. Let the population proportion be defined as $\theta = \frac{1}{N}\sum_{i\in U} y_i$, treating y as a qualitative sensitive variable - *"Breaking the traffic rules"*. For the above study, $\theta = 0.606838$. The innocuous character x is taken here as *"Interested in painting"*. The size measure variable z- *"Weekly expenditure"* is used to draw samples in varying probability sampling scheme.

In order to study the competitiveness concerning the proposed RRTs, the simulation study is performed for different sample sizes and the samples are drawn by Lahiri- Midzuno- Sen [1951, 1952, 1953] sampling strategy, where the first unit is selected with the probability $p_i^* = \frac{z_i}{\sum_U z_i}$ and the remaining units are selected by SRS without replacement from the remaining units in the population after the first draw.

Since the variable y represents the sensitive feature A, it is not directly assessable and is estimated for each respondent through an unbiased estimator defined in eq.1. Then, employing eq. 3 and eq. 4, we get an unbiased estimate for the population proportion and unbiased variance estimate respectively. Here, the 1st order and 2nd order inclusion probabilities for Lahiri- Midzuno- Sen scheme are $\pi_i = p_i^* + \frac{(1-p_i^*)(n-1)}{N-1}$ and $\pi_{ij} = \frac{(n-1)(N-n)(p_i^*+p_j^*)+(n-1)(n-2)}{(N-1)(N-2)}$ respectively.

To judge the efficacy of the RR models, different parametric combinations (p_1, p_2) are taken in Table 5.1, considering 1000 replications of samples for each sample size. Efficacy of the proposed RRTs for different sample sizes are judged by the Average Coverage Probabilities (ACP), the Average Coefficient of Variation (ACV) and the Average Length (AL) of the 95% Confidence Intervals (CI) based on $e_{HT} \pm 1.96\sqrt{v(e_{HT})}$. The point estimator will be judged well if the ACV, the average over 1000 replications of estimated coefficient of variations$\left(CV = 100 \times \frac{\sqrt{v(e_{HT,RR})}}{e_{HT,RR}}\right)$, has a small magnitude, preferably less than 10% or at most 30%. The percentage of cases for which 95% CI covers the true value of the parameter is called ACP. ACP values close to 95% will be preferred. AL is defined as $2 \times 1.96\sqrt{[\![v(e]\!]_{HT}}$. Smaller ACV, AL values along with the ACP value close to 95% are preferred. In addition, absolute relative bias (ARB) of an

unbiased estimator and average variance estimate (AVE) are calculated as $\left|\frac{\bar{e}-\theta}{\theta}\right|$ and $\frac{1}{1000}\sum_{k=1}^{1000} v(e_k)$ respectively, where $\bar{e} = \frac{1}{1000}\sum_{k=1}^{1000} e_k$ is the average of 1000 estimates of θ.

Figures 5.1-5.3, based on Table 5.1, represent the performance of proposed RRTs for different sample sizes. The values of ACV, ACP, AL, ARB and AVE for different parametric combinations are shown in the same graph. To do this, the values of AL, ARB and AVE from Table 5.1 are taken as $100 \times AL$, $1000 \times ARB$ "and" $1000 \times AVE$. The vertical axes of the graphs indicate the values and the horizontal axes indicate different parametric combinations (p_1, p_2) of the RRTs.

As shown in Table 5.1,

i) ACP values are greater than 95%.

ii) ACV and AL values are decreasing as the sample size increases. If p_1 and p_2 are close to each other, ACV and AL values are relatively high.

For example:

if $(p_1, p_2) = (0.4, 0.6)$ and $(0.4, 0.2)$, ACV values are beyond the acceptable range.

iii) Considering the sample size $n = 25$, the parametric combinations (0.4, 0.7) and (0.57, 0.79) are equally competitive and perform moderately as their ACV, AL and AVE values are much lower than others. Figure 5.1 sheds light on this fact.

iv) The parametric combinations (0.4, 0.7) and (0.57, 0.79) perform well in terms of AVE, ACV, ACP, ARB and AL for both sample sizes 30 and 35. Figures 5.2 and 5.3 highlight this finding.

Table 5.1: ACV, ACP, AL, and ARB for the proposed RRTs using fivetypes of cards $(unknown\ \theta = 0.606838)$

p_1	p_2	n	$\bar{e} = \frac{1}{1000}\sum_{k=1}^{1000} e_k$	$AVE = \frac{1}{1000}\sum_{k=1}^{1000} v(e_k)$	ACV	ACP	AL	ARB
0.4	0.6	25	0.61445	0.04718	41.07467	98.7	0.83896	0.01255
0.4	0.6	30	0.61157	0.03981	35.25056	99	0.77313	0.00781
0.4	0.6	35	0.62353	0.03267	30.60223	98.7	0.70227	0.02751
0.4	0.2	25	0.68875	0.08238	46.29767	96.3	1.10524	0.13498
0.4	0.2	30	0.65915	0.07464	43.47351	97.6	1.05719	0.08620
0.4	0.2	35	0.65619	0.06093	42.15731	97.4	0.95809	0.08133
0.4	0.7	25	0.62845	0.01998	23.3216	97.4	0.54935	0.03561
0.4	0.7	30	0.61883	0.01730	21.89433	98	0.51232	0.01976
0.4	0.7	35	0.62145	0.01408	19.54249	98.3	0.46274	0.02408
0.57	0.79	25	0.62094	0.01941	23.46384	96.3	0.54168	0.02324
0.57	0.79	30	0.62819	0.01629	21.01769	96.1	0.49702	0.03519
0.57	0.79	35	0.62929	0.01336	18.78795	98.2	0.45075	0.03699
0.57	0.3	25	0.65181	0.03375	30.09188	96.6	0.71012	0.07411
0.57	0.3	30	0.65319	0.02885	27.31937	96.7	0.65903	0.07639
0.57	0.3	35	0.65078	0.02344	24.54986	96.8	0.59563	0.07241

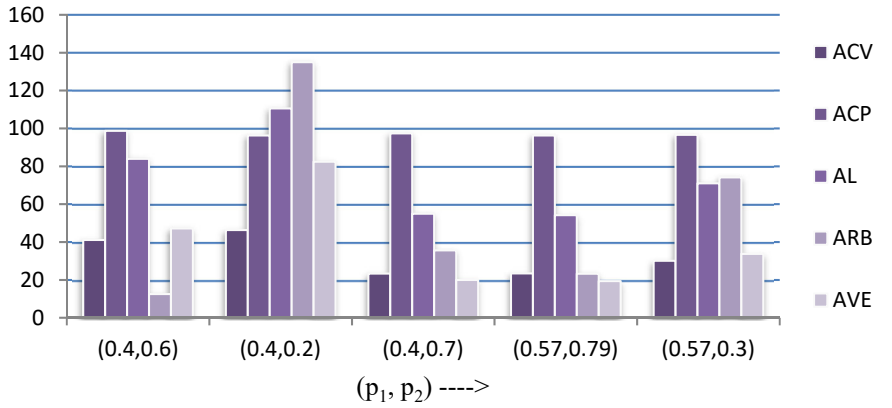

Figure 5.1: Performances of the proposed RRTs for the sample size n=25

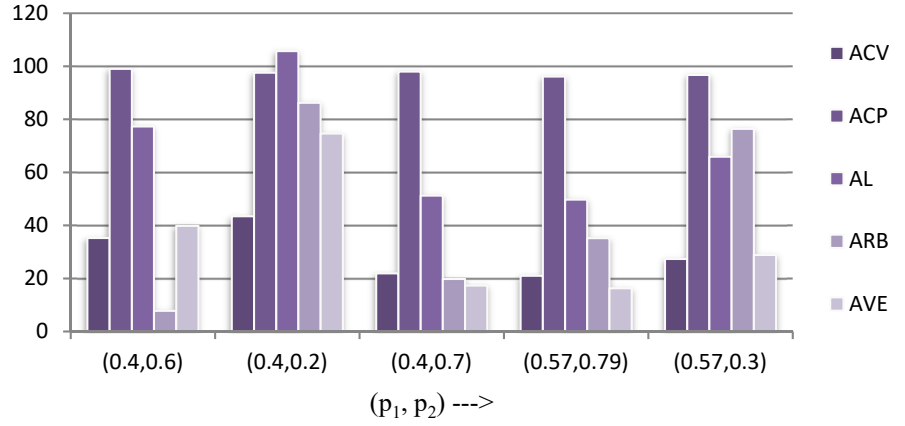

Figure 5.2: Performances of the proposed RRTs for the sample size n=30

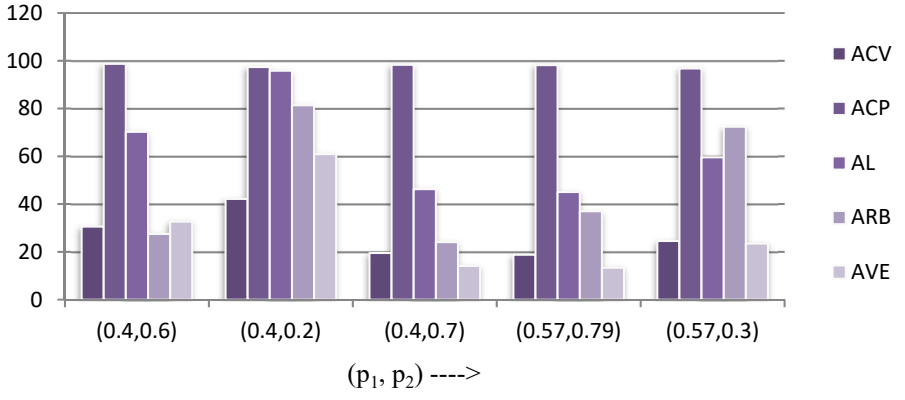

Figure 5.3: Performances of the proposed RRTs for the sample size n=35

Table 5.2 demonstrates how well the privacy of sensitive features may be protected for the proposed models. For this purpose, the response-specific jeopardy measures $J_i(R, R')$ are computed taking different combinations of $(p_1, p_2, w_2, w_3, w_4)$. The overall measure of jeopardy $(\tilde{J}_i)$ is shown in the last column of the mentioned table.
As noted in Section 4,

i) $J_i(R, R') \to 1$ if $p_1 \to p_2$ and the proportion of *"I possess A"* cards tends to the proportion of *"I possess A^c"* cards, which also ensures that the overall measure of jeopardy tends to 1.

In Table 5.2, we have tried to show such a condition taking the combinations of $(p_1, p_2, w_2, w_3, w_4)$values as (0.33,0.327, 0.49,0.2,0.3) and (0.2,0.22,0.27,0.2,0.4).

ii) In both RR devices if the proportion of "Yes" cards tends to the proportion of "No" cards, the measure of jeopardy $\tilde{J}_i$ will tend to 1.

In Table 5.2, the following parameters, satisfying the above conditions, may be taken as follows:

$(p_1, p_2, w_2, w_3, w_4)$: (0.4, 0.45, 0.2,0.3,0.25) , (0.2,0.3,0.29,0.2,0.27) and (0.4,0.6,0.2,0.3,0.25).

In Table 5.3 and Table 5.4, we have shown the response-specific jeopardy measure along with the overall measure of jeopardy following the suggestion in Chaudhuri et al. (2009) for Warner's (1965) RRT and Greenberg et al.'s (1969) RRT. Here, the arithmetic mean $(\bar{J}_i)$ of all the response-specific jeopardy measure is considered as the overall measure of jeopardy. We refer to Chaudhuri et al. (2009) for detailed derivation of the measures. The results in Table 5.2 can be compared easily with Table 5.4. For better comparison, we have taken the same $p_1$and p_2 values which represent the proportions of "I possess A" cards for proposed RRTs (see Section 3) and Greenberg et al.'s RRT. From there, we may conclude that the proposed RRTs perform better than the Greenberg et al.'s RRT in terms of protecting privacy.

Table 5.2: Protection of Privacy for Proposed RRTs

p_1	p_2	w_2	w_3	w_4	$J_i(1,1)$	$J_i(0,0)$	$J_i(1,0)$	$J_i(0,1)$	$\tilde{J}_i$
0.4	0.6	0.2	0.3	0.4	3.7119	0.1776	0.4795	1.375	0.8120
0.4	0.6	0.2	0.3	0.2	4.7619	0.2406	0.6349	1.8045	1.0704
0.4	0.45	0.2	0.3	0.25	2.9593	0.3379	0.8892	1.1245	1
0.33	0.38	0.49	0.2	0.3	1.1270	0.8476	0.8528	1.1201	0.9774
0.33	0.327	0.49	0.2	0.3	0.9984	1.0022	1.0085	0.9922	1.0003
0.57	0.69	0.2	0.45	0.2	7.8635	0.1176	0.6579	1.4056	0.9617
0.6	0.5	0.2	0.45	0.2	4.9100	0.1905	1.2647	0.7395	0.9671
0.2	0.22	0.27	0.2	0.4	0.9905	1.0141	0.9578	1.0488	1.0023
0.2	0.3	0.29	0.2	0.27	1.1201	0.8892	0.7962	1.2509	0.998
0.2	0.3	0.4	0.2	0.3	0.8598	1.2228	0.8006	1.3131	1.0254
0.4	0.6	0.2	0.3	0.25	4.4340	0.2255	0.5935	1.6849	1

Table 5.3: Protection of Privacy for Warner's RRT

p_1	$J_i(1) = \frac{p_1}{1-p_1}$	$J_i(0) = \frac{1-p_1}{p_1}$	$\bar{J}_t$
0.2	0.25	0.4	2.125
0.33	0.49254	2.0303	1.26142
0.4	0.66667	1.5	1.08333
0.51	1.04081	0.96078	1.0008
0.57	1.32558	0.75439	1.03998
0.6	1.5	0.66667	1.08333
0.69	2.22581	0.44927	1.33754

Table 5.4: Protection of Privacy for Greenberg et al.'s RRT

p_1	p_2	$J_i(1,1) = \frac{p_1 p_2}{(1-p_1)(1-p_2)}$	$J_i(0,0) = \frac{(1-p_1)(1-p_2)}{p_1 p_2}$	$J_i(1,0) = \frac{p_1(1-p_2)}{p_2(1-p_1)}$	$J_i(0,1) = \frac{p_2(1-p_1)}{p_1(1-p_2)}$	$\bar{J}_t$
0.4	0.6	1	1	0.4444	2.25	1.1736
0.4	0.45	0.5455	1.8333	0.8149	1.2273	1.1052
0.33	0.38	0.3019	3.3126	0.8036	1.2444	1.4156
0.57	0.69	2.9505	0.3389	0.5955	1.6791	1.3910
0.6	0.5	1.5	0.6667	1.5	0.6667	1.0833
0.2	0.3	0.1071	9.3333	0.5833	1.7143	2.9345

6. Concluding Remarks

In this work, we have attempted to introduce two proposed methods permitting five questions to the respondents. Model 2, i.e. Respondent-specific RRT, is an extension of the proposed Model 1. In the proposed Model 2, respondents are allowed to build their own RR devices. It is anticipated that the participation of respondents in the RR survey will be better than other existing RRTs. Our simulation study gives us satisfactory results in terms of ACP, ACV and AL values. We have also calculated protection of privacy measure of the proposed RRTs, which is close to 1. The findings described in this study will stimulate researchers and survey practitioners to apply the response-specific RRT in real surveys. Respondents will co-operate freely in the survey methods as they are building their own RR devices.

Acknowledgement

The authors gratefully acknowledge the support received from two referees, which enabled them to produce this revised and improved version out of the original submission.

References

Anderson, H., (1975b). Efficiency versus Protection in a General RR model, *Technical Report 10, University of Lund,* Lund, Sweden.

Anderson, H., (1975c). Efficiency versus Protection in RR Designs. *Mimeo notes, University of Lund,* Lund, Sweden.

Anderson, H., (1975a). Efficiency versus Protection in the RR for Estimating Proportions. *Technical Report 9, University of Lund, Lund, Sweden*.

Arnab, R., Mothupi, T., (2015). Randomized Response Techniques: A Case Study of the Risky Behaviors' of Students of a Certain University. *Model Assisted Statistics and Applications*, Vol. 10, pp. 421–430.

Barabesi, L., Diana, G., Perri, P. F., (2013). Design-based distribution function estimation for stigmatized population. *Metrika*, Vol. 76, pp. 919–935.

Boruch, R. F., (1972). Relations Among Statistical Methods for Assuring Confidentiality of Social Research Data. *Social Science Research*, Vol. 1, pp. 403–414.

Chaudhuri, A., (2011). *Randomized Response and Indirect Questioning Techniques in Surveys*. Boca Raton: CRC Press.

Chaudhuri, A., Christofides, T. C., Rao, C. R., (2016). Handbook of Statistics, Data Gathering, *Analysis and Protection of Privacy Through Randomized Response Techniques: Qualitative and Quantitative Human Traits*, Vol. 34, NL: Elsevier.

Chaudhuri, A., Christofides, T. C., Saha, A., (2009). Protection of Privacy in Efficient Application of Randomized Response Techniques. *Statistical Methods and Applications*, Vol. 18, pp. 389–418.

Chong, A. C., Chu, A. M., So, M. K., Chung, R. S., (2019). Asking Sensitive Questions Using the Randomized Response Approach in Public Health Research: An Empirical Study on the Factors of Illegal Waste Disposal. *International Journal of Environmental Research and Public Health*, Vol. 16, pp. 970.

Diana, G., Perii, P. F., (2013). Randomized Response Surveys: A note on some privacy protection measures. *Model Assisted Statistics and Applications*, Vol. 8, pp. 19–28.

Fox, J., Veen, D., Klotzke, K., (2018). Generalized linear mixed models for randomized responses. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences,* Vol. 15, pp. 1–18.

Greenberg, B. G., Abul-Ela, A. L., Simmons, W. R.,Horvitz, D. G., (1969). The unrelated question randomized response model: theoretical framework. *Journal of American Statistical Association*, Vol. 64, pp. 520–539 .

Horvitz, D. G., Thompson, D. J., (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, Vol. 47, pp. 663–685 .

Hout, A. V., Heijden, P. G., Gilchrist, R., (2007). The logistic regression model with response variables subject to randomized response. *Computational Statistics & Data Analysis,* Vol. 51, pp. 6060–6069.

Kuk, A. Y., (1990). Asking Sensitive Questions Indirectly. *Biometrika*, Vol. 77, pp. 436–438.

Lahiri, D. B., (1951). A method of sample selection providing unbiased ratio estimates. *Bulletin of International Statistical Institute*, Vol. 3, pp. 133–140.

Lanke, J., (1975). On the choice of the unrelated question in Simmons' version of randomized response. *Journal of American Statistical Association*, Vol. 70, pp. 80–83 .

Lanke, J., (1976). On the degree of protection in randomized interviews. *International Statistical Review*, Vol. 44, pp. 197–203.

Leysieffer, R. W., Warner, S. L., (1976). Respondent jeopardy and optimal designs in randomized response models. *Journal of American Statistical Association*, Vol. 71, pp. 649–656 .

Midzuno, H., (1952). On the sampling system with probability proportional to the sum of the sizes. *Annals of Institute of Statistical Mathematics*, Vol. 3, pp. 99–107.

Pal, S., Chaudhuri, A., Patra, D., (2020). How privacy may be protected in Optional Randomized Response Surveys. *Statistics in Transition*, Vol. 21, pp. 61–87.

Sen, A. R., (1953). On the estimator of the variance in sampling with varying probabilities. *Journal of Indian Society of Agricultural Statistics*, Vol. 5, pp. 119–127.

Van der Heijden, P. G., Van Gils, G., Bouts, J., Hox, J., (2000). A Comparison of Randomized Response, Computer-Assisted Self Interview, and Face to Face Direct Questioning: Eliciting Sensitive Information in the Context of Welfare and Unemployment Benefit. *Sociological Research & Methods*, Vol. 28, pp. 505–537 .

Warner, S. L., (1965). Randomized response: a survey technique for eliminating evasive answer bias. *Journal of American Statistical Association*, Vol. 60, pp. 63–69.