I^2SDS
The Institute for Integrating Statistics in Decision Sciences

Technical Report TR-2007-12
September 17, 2007

On Randomized Response Surveys of Binary Characteristics Using Polychotomous Response Variables

Tapan K. Nayak
Department of Statistics
The George Washington University, USA

Samson A. Adeshiyan
Department of Statistics
The George Washington University, USA

On Randomized Response Surveys of Binary Characteristics Using Polychotomous Response Variables

Tapan K. Nayak and Samson A. Adeshiyan

Department of Statistics

George Washington University

Washington, DC 20052, USA.

Abstract

The randomized response (RR) procedures for estimating the proportion (π) of a population belonging to a sensitive or stigmatized group ask each respondent to report a response by randomly transforming his/her true attribute into one of several response categories. In this paper, we present a common framework for discussing various RR surveys of dichotomous populations with polychotomous responses. The unified approach is focused on the substantive issues relating to respondents' privacy and statistical efficiency and is helpful for fair comparison of various procedures. We describe a general technique for obtaining unbiased estimators of π based on RR data, from unbiased estimators of π based on open surveys. The technique works well for any sampling design $p(s)$ and also for variance estimation. We develop an approach for comparing RR procedures, taking both respondents' protection and statistical efficiency into account. For any given RR procedure with three or

more response categories, we present a method for designing an RR procedure with a binary response variable which provides the same respondents' protection and at least as much statistical information. This result suggests that RR surveys of dichotomous populations should use only binary response variables.

Key words and Phrases: Design unbiasedness, Fisher information, respondents' protection, sampling design, variance estimator.

1. Introduction.

In most surveys, the individuals selected in the sample are asked to answer direct questions relating to the survey variables. However, for questions on sensitive or stigmatizing characteristics such as criminal history, tax evasion, drug abuse, gambling and abortion, many respondents are unwilling to give honest answers, if at all they respond in the first place. The refusals and false answers lead to biased and unreliable estimates. The main reason for lack of respondents' cooperation is the lack of privacy. To increase truthful respondent participation, Warner (1965) proposed the first randomized response (RR) procedure for a binary characteristic, which we briefly review next. Consider a dichotomous population where each person belongs either to a sensitive group A or to its complement A^c. The objective is to estimate the true proportion (π) of the population that belongs to group A. In Warner's (1965) method, each interviewee first selects one of the two questions:

Q_1: Do you belong to A?

Q_2: Do you belong to A^c?

with respective probabilities p and $(1-p)$, by performing a random experiment, unobserved by the interviewer. The respondent then truthfully replies "Yes" or "No" to the selected question without disclosing the question and thereby protecting his/her privacy. The probabilities p and $(1-p)$ are known and are embedded in the randomization mechanism. In this scheme, the probability of the "Yes" response is:

$$P_W(\text{Yes}) = \lambda_W = \pi p + (1-\pi)(1-p) = (1-p) + (2p-1)\pi.$$

Let n denote the sample size and X denote the number of "Yes" responses. Considering $p \neq 0.5$ and simple random sampling with replacement (SRSWR), where $X \sim b(n, \lambda_W)$, Warner

(1965) proposed the following method of moments estimator of π:

$$\hat{\pi}_W = \frac{\hat{\lambda}_W - (1-p)}{2p-1},$$

where $\hat{\lambda_W} = X/n$. The variance of $\hat{\pi}_W$ is

$$Var(\hat{\pi}_W) = \frac{\lambda_W(1-\lambda_W)}{n(2p-1)^2} = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2}. \tag{1.1}$$

The last two terms of (1.1) represent, respectively, the variance of the minimum variance unbiased estimator of π from an open (or direct) survey and the additional variance due to randomization. Both the variance of $\hat{\pi}_W$ and the degree of respondents' privacy depend on the value of p.

Greenberg et al. (1969) discussed a related procedure, called Simmons' unrelated question method, in which the question Q_2 in Warner's method is replaced by an unrelated nonsensitive question:

Q_3: Do you belong to B?

An example of an unrelated question is: were you born in the month of June? The probabilities of "yes" (Y) and "no" (N) responses to Q_3 may be known, in which case, a method of moments estimator ($\hat{\pi}_U$) of π can be derived easily. Greenberg et al. (1969) showed that $Var(\hat{\pi}_U)$ is smaller than $Var(\hat{\pi}_W)$ when the probability (p) of asking the direct question Q_1 is the same in the two methods. As noted by Leysieffer and Warner (1976) and Fligner et al. (1977), that comparison is not fair because the two procedures, with a common value of p, offer different degrees of privacy to the respondents. For fair comparison, the two procedures should be required to offer equal respondents' protection. Several other RR methods have been proposed and investigated in the literature, e.g., Chaudhuri and Mukerjee (1988), Kuk (1990), Mangat and Singh (1990), Mangat (1994) and Kim and Warde (2004). However, some of the efficiency comparisons, e.g.,

Greenberg et al. (1969), Mangat and Singh (1990) and Mangat (1994), are flawed as they do not hold respondents' protection at the same level.

Various RR methods can be found in the literature, but each RR procedure has usually been discussed using features (parameters) that are specific to its randomization mechanism. Often the randomization mechanisms of two procedures share a common element, but with different effect on privacy and efficiency of the two procedures. For example, the question Q_1 is common to Warner's and Simmons' procedures, but the probability (p) of asking Q_1 affects respondents' protection and statistical efficiency differently for the two procedures; see Fligner et al. (1977) for a more detailed discussion and some numerical illustrations. Some unfair comparisons stemmed from considering two procedures with a common randomization parameter, but with disparate impact on respondents' protection, and then comparing variances of the estimators proposed under the two procedures. We believe, misleading comparisons could be avoided by discussing various RR procedures within a common framework. A general framework is also important for identifying and placing the substantive logical issues at the forefront. For binary response RR surveys of dichotomous populations, Nayak (1994) proposed a unified framework, which we briefly discuss below.

Let Y be an indicator of the sensitive characteristic, viz., $Y = 1$ if the respondent belongs to the sensitive group A and $Y = 0$ otherwise. Let $Z = 0$ and $Z = 1$ label the two response categories. For example, in Warner's and Simmons' procedures, the "Yes" and "No" responses may be recorded as $Z = 1$ and $Z = 0$, respectively. Let, a and b denote $P(Z = 1|Y = 1)$ and $P(Z = 1|Y = 0)$, respectively. Then, the posterior probabilities of $Y = 1$, which determine the

level of respondents' privacy, are:

$$P(Y=1|Z=1) = \frac{a\pi}{a\pi + b(1-\pi)}$$
$$P(Y=1|Z=0) = \frac{(1-a)\pi}{(1-a)\pi + (1-b)(1-\pi)}.$$

Let n denote the sample size and X denote the number of respondents reporting $Z=1$. Then, under the common assumptions of simple random sampling with replacement (SRSWR) and truthful answering, $X \sim b(n,\theta)$, where $\theta = a\pi + b(1-\pi)$. It can be seen that if $a \neq b$, the uniformly minimum variance unbiased estimator of π, based on X, is

$$\hat{\pi} = (\frac{X}{n} - b)/(a-b)$$

and its variance is $V(\hat{\pi}) = \theta(1-\theta)/[n(a-b)^2]$. If a and b are known and are determined only by the randomization mechanism, as is the case for most binary response RR procedures, all statistical properties, including protection of privacy and accuracy of statistical inferences, depend on the randomization step only through the values of a and b. So, such procedures can be characterized by a and b, taking them as the RR design parameters. Thus, a unified approach ensues from discussing various binary response RR procedures in terms of their design parameters a and b.

Remark 1. Any one-to-one transformation of (a,b) can also be used as the RR design parameters for developing a unified framework. In particular, Nayak (1994) used $P(Yes|A)$ and $P(No|A^c)$ as the RR design parameters, which correspond to our a and $1-b$ if $Z=1$ and $Z=0$ represent the "Yes" and "No" responses, respectively. Leysieffer and Warner (1976) expressed respondents' protection and $Var(\hat{\pi})$ in terms of $u = a/b$ and $v = (1-b)/(1-a)$. As the transformation $\{a,b\} \to \{u,v\}$ is one-to-one, a unified framework can also be developed in terms of u and v.

Remark 2. As it was noted in Nayak (1994), the interchanging of the two responses "Yes" and "No" (or equivalently $Z = 1$ and $Z = 0$) does not alter any statistical property of a procedure, which implies that for any $0 \leq a, b \leq 1$, the two RR procedures with RR design parameters (a, b) and $(1 - a, 1 - b)$, respectively, are equivalent. For unique representation, we may impose the restriction $a > b$ and take $\{(a, b) : 0 \leq a, b \leq 1, a > b\}$ as the RR design space. In this framework, Nayak (1994) showed that respondents' protection and statistical efficiency do not necessarily move in opposite directions and an RR design (a, b) is admissible if and only if $a = 1$.

Remark 3. The general framework presented above covers all binary response RR procedure for which the randomization probabilities a and b are known. It does not cover Simmons' two sample procedure, where the probability of the "Yes" answer to Q_3 is unknown and the sampled individuals are divided into two groups to receive the questions Q_1 and Q_3 with different but known probabilities (Greenberg et al., 1969).

Some RR procedures proposed in the literature use polychotomous responses (e.g., Leysieffer and Warner, 1976; Kuk, 1990; Christofides, 2003). In Chow's procedure, discussed in Leysieffer and Warner (1976), each respondent selects k balls at random, without replacement and unobserved by the interviewer from an urn containing L red and M blue balls, where L and M are known and $k \leq \min\{L, M\}$. The respondent then reports the number of red balls if he/she belongs to the sensitive group A; otherwise he/she reports the number of blue balls. Christofides (2003) proposed a similar, albeit more general, procedure where each person in the sample is provided with a device which produces the integers $1, \cdots, k$ with known probabilities $p_1, \cdots, p_k$, respectively. Each respondent uses the device, in the absence of the interviewer, to produce one integer J and then reports the value of $(k + 1 - J)$ if he/she belongs to A; otherwise, he/she

reports the value of J. In Kuk's (1990) repeated trials design, each respondent is given two decks of cards. Both decks comprise of cards of two colors, say red and blue, but with different proportions. A respondent selects k cards at random and with replacement from deck 1(2) if he/she belongs to $A(A^c)$ and reports the number of red cards selected. In all of these procedures, the response variable Z is integer valued and the probabilities $\{P(Z = z|A)\}$ and $\{P(Z = z|A^c)\}$ are known and specified by the randomization device.

The main goals of this paper are to present a unified framework for RR surveys of dichotomous populations with polychotomous response variables, discuss unbiased estimation under general sampling designs and compare RR surveys, paying attention to both respondents' protection and statistical efficiency. In Section 2, we lay out a unified framework and express privacy measures and some basic statistical entities in terms of the randomization parameters. We hope the proposed framework will be helpful for thinking in a principled way about privacy and statistical efficiency. Most papers present estimators for SRSWR, but many surveys employ unequal probability sampling, e.g., stratified and multi-stage sampling. In Section 3, we discuss unbiased estimation of π and variances of estimators under a general sampling design $p(s)$. Following Padmawar and Vijayan (2000) and Chaudhuri (2001, 2004), we present a technique for amending a linear unbiased estimator under an open survey to obtain an unbiased estimator for an RR survey. We also uncover and discuss an arbitrariness inherent in that approach. In Section 4, we compare RR surveys taking both respondents' protection and statistical information into account. We find that use of polychotomous responses is not really helpful for sampling dichotomous populations. Specifically, given any RR procedure with a polychotomous response variable, we can devise a better RR procedure using a dichotomous response variable. Section 5 presents some concluding remarks.

2. A Unified Framework.

Let us now consider RR surveys of dichotomous populations using polychotomous response variables. As before, let $Y = 1$ if the respondent belongs to the sensitive group A and $Y = 0$ if the respondent belongs to A^c and let $\pi = P(Y = 1)$ be the unknown parameter of interest. We shall denote the response variable by Z, the number of response categories by $k(k \geq 2)$ and the possible responses by $c_1, \cdots, c_k$, satisfying $c_i \neq c_j$ for $i \neq j$. Further, let $\alpha_i = P(Z = c_i|Y = 1), \beta_i = P(Z = c_i|Y = 0)$ for $i = 1, \cdots, k$, $\vec{\alpha} = (\alpha_1, ..., \alpha_k)$ and $\vec{\beta} = (\beta_1, ..., \beta_k)$. For uniqueness of k, we shall require that $\min(\alpha_i, \beta_i) > 0, i = 1, \cdots, k$. We shall call a RR survey with k response categories a $(2 \to k)$ RR survey. The binary response surveys, as considered in Nayak (1994), correspond to $k = 2$. We shall consider all $(2 \to k)$ RR surveys with known $\vec{\alpha}$ and $\vec{\beta}$, noting that for protecting respondent's privacy it is not necessary to use a randomization device for which $\vec{\alpha}$ and $\vec{\beta}$ are unknown.

Let θ_i denote $P(Z = c_i)$, i.e., $\theta_i = \alpha_j\pi + \beta_j(1 - \pi)$. Then, the posterior probabilities which determine the level of respondents' privacy are:

$$P(A|Z = c_j) = \frac{\alpha_j\pi}{\alpha_j\pi + \beta_j(1 - \pi)} = \frac{\pi}{\pi + (\beta_j/\alpha_j)(1 - \pi)}, \quad j = 1, \cdots, k. \tag{2.1}$$

For $i = 1, \cdots, k$, let X_i denote the observed frequency of the response c_i. Then, under SRSWR, $(X_1, \cdots, X_k) \sim mult(n; \theta_1, \cdots, \theta_k)$ with probability mass function

$$f_\pi(x_1, \cdots, x_k) = P(X_1 = x_1, \cdots, X_k = x_k) = \frac{n!}{x_1! \cdots x_k!}\theta_1^{x_1} \cdots \theta_k^{x_k}.$$

The maximum likelihood estimate (MLE) of π, based on $(X_1, \cdots, X_k)$, is the solution of

$$\frac{\partial}{\partial\pi} \ln f_\pi(x_1, \cdots, x_k) = 0 \quad \text{or} \quad \sum_{i=1}^{k} \frac{x_i(\alpha_i - \beta_i)}{\alpha_i\pi + \beta_i(1 - \pi)} = 0,$$

provided that the solution is in $[0, 1]$. In the case of sampling from an infinite population, the Fisher information in a single response is

$$i(\pi) = \sum_{i=1}^{k} \frac{(\alpha_i - \beta_i)^2}{\alpha_i \pi + \beta_i (1 - \pi)}$$

and the asymptotic distribution of the MLE ($\hat{\pi}_{ML}$) is normal:

$$\sqrt{n}(\hat{\pi}_{ML} - \pi) \xrightarrow{L} N(0, i(\pi)) \quad \text{as } n \to \infty,$$

which can be used to construct large sample confidence intervals for π. Note that the posterior probabilities in (2.1), the distribution of $(X_1, \cdots, X_k)$ and the Fisher's information depend on the randomization mechanism only through $\vec{\alpha}$ and $\vec{\beta}$. Thus, all $(2 \to k)$ RR procedures can be characterized by the values of $\vec{\alpha}$ and $\vec{\beta}$. This implies that for designing a $(2 \to k)$ RR survey we should first determine the values of $\vec{\alpha}$ and $\vec{\beta}$ and then devise a mechanism for implementing them. For a unified approach, we suggest to take $(\vec{\alpha}, \vec{\beta})$ as the RR design parameters and discuss and examine all $(2 \to k)$ RR procedures through them.

Remark 4. The ordering of the k response categories should have no bearing on the substantive properties of a $(2 \to k)$ RR design. A $(2 \to k)$ RR design essentially remains unchanged under any permutation of the response categories and corresponding permutations of the components of $\vec{\alpha}$ and $\vec{\beta}$. Thus, for any $(\vec{\alpha}, \vec{\beta}) = (\alpha_1, \cdots, \alpha_k, \beta_1, \cdots, \beta_k)$ and any permutation $(i_1, \cdots i_k)$ of $(1, \cdots, k)$, the two $(2 \to k)$ RR designs with randomization probabilities $(\vec{\alpha}, \vec{\beta})$ and $(\alpha_{i_1}, \cdots, \alpha_{i_k}, \beta_{i_1}, \cdots, \beta_{i_k})$, respectively, are equivalent. This implies that any $(2 \to k)$ RR design can be characterized by many different sets of values of the RR design parameters $(\vec{\alpha}, \vec{\beta})$, generated by permutations of the response categories. So, for unique characterization of a $(2 \to k)$ RR design by $(\vec{\alpha}, \vec{\beta})$, we need a convention for ordering the response categories. One possibility is to order the categories first by the values of $\{\alpha_i\}$ and then by the β values, i.e.,

require that $\alpha_1 \geq \cdots \geq \alpha_k$ and if $\alpha_j = \cdots = \alpha_{j+m}$ for any j and m, then $\beta_j \geq \cdots \geq \beta_{j+m}$. Noting that the posterior probability $P(A|Z = c_j)$ in (2.1) depends on $\vec{\alpha}$ and $\vec{\beta}$ only through the ratio (α_i/β_i) and $P(A|Z = c_j)$ is an increasing function of (α_i/β_i), we believe a more meaningful approach would be to order the categories first in decreasing order of magnitude of (α_i/β_i), i.e., in decreasing order of the probability of being classified in the sensitive group A, and then by decreasing order of magnitude of $\{\alpha_i\}$. Thus, to make $(\vec{\alpha}, \vec{\beta})$ unique we suggest to require $(\alpha_1/\beta_1) \geq \cdots \geq (\alpha_k/\beta_k)$ and if $(\alpha_j/\beta_j) = \cdots = (\alpha_{j+m}/\beta_{j+m})$ for any j and m, then $\alpha_j \geq \cdots \geq \alpha_{j+m}$.

3. Estimation.

While most authors discussed statistical analyses of RR data assuming random sampling from an infinite population or SRSWR, practical surveys often involve complex survey designs and many variables, only a few of which may be sensitive. Thus, it is important to derive estimators based on RR data and unequal probability sampling. For a quantitative response variable, Padmawar and Vijayan (2000) discussed linear unbiased estimation of a finite population total based on RR data obtained under a general sampling design. Analogously, Chaudhuri (2001, 2004) presented linear unbiased estimators of a population proportion (π) based on certain RR procedure including the Christofides' (2003) procedure. Specifically, they showed how a linear unbiased estimator based on an open survey can be modified to obtain an unbiased estimator under an RR survey. They also discussed unbiased estimation of the variance of the estimators from RR survey data. In this section, we first develop similar results for a general $(2 \to k)$ RR procedure and then discuss an arbitrariness of the approach.

3.1. Estimation of π.

Consider a finite population of N units, labeled $i = 1, \cdots, N$, and let Y_i denote the value of Y (an indicator of the sensitive variable) for unit i. Let Z_i denote the response of unit i and suppose the sample is selected using a non-informative sampling design $p(s)$. So, the data can be represented as $\{(i, Z_i); i \in s\}$, where s is a subset of $\{1, \cdots, N\}$, and our goal is to estimate $\pi = (\sum_{i=1}^{N} Y_i)/N$. Note that while Y_i are fixed, Z_i are random variables, and estimation of π is equivalent to estimation of the population total $T(Y) = \sum_{i=1}^{N} Y_i$.

Suppose

$$e(s, y) = w_{s0} + \sum_{i \in s} w_{si} Y_i \tag{3.1}$$

is a linear unbiased estimator $T(Y)$, i.e.,

$$E_p[e(s, y)] = \sum_{s} e(s, y)p(s) = \sum_{i=1}^{N} Y_i \quad \text{for all } Y_1, \cdots, Y_N$$

or equivalently,

$$\sum_{s} w_{s0} p(s) = 0 \quad \text{and} \quad \sum_{s \ni i} w_{si} p(s) = 1, \quad i = 1, \cdots, N.$$

To extend Chaudhuri's (2001, 2004) results, in this section we shall require $c_1, \cdots, c_k$ to be real numbers. Then, since $P(Z = c_j|Y = 1) = \alpha_j$ and $P(Z = c_j|Y = 0) = \beta_j$, we have

$$E[Z|Y = 1] = \sum_{j=1}^{k} \alpha_j c_j \quad \text{and} \quad E[Z|Y = 0] = \sum_{j=1}^{k} \beta_j c_j$$

or

$$E[Z|Y] = \sum_{j=1}^{k} \beta_j c_j + [\sum_{j=1}^{k} (\alpha_j - \beta_j) c_j] Y = d_1 + d_2 Y, \quad \text{say,}$$

where $d_1 = \sum_{j=1}^{k} \beta_j c_j$ and $d_2 = \sum_{j=1}^{k} (\alpha_j - \beta_j) c_j$. So, if $d_2 \neq 0$, i.e., $(\vec{\alpha} - \vec{\beta})$ is not orthogonal to $\vec{c} = (c_1, \cdots, c_k)$, letting $U = (Z - d_1)/d_2$, it follows that $E_R(U) = E(U|Y) = Y$, where E_R

denotes expectation with respect to the randomization mechanism. Let

$$e^*(s,z) = w_{s0} + \sum_{i\in s} w_{si}U_i = w^*_{s0} + \sum_{i\in s} w^*_{si}Z_i, \tag{3.2}$$

where $w^*_{s0} = w_{s0} - (d_1/d_2)\sum_{i\in s} w_{si}$ and $w^*_{si} = w_{si}/d_2$. Then, it can be seen that

$$E[e^*(s,z)] = E_pE_R[w_{s0} + \sum_{i\in s} w_{si}U_i] = E_p[e(s,y)] = T(Y).$$

Thus, we have the following:

Theorem 1. For any given sampling design $p(s)$, if $e(s,y)$ in (3.1) is a linear design unbiased estimator of the population total $T(Y) = \sum_{i=1}^N Y_i$ based on the open survey, then the estimator $e^*(s,z)$ in (3.2) is a linear design unbiased estimator of $T(Y)$ based on the RR survey with RR design $(\vec{\alpha}, \vec{\beta})$ and sampling design $p(s)$.

Conversely, from any given linear unbiased estimator for an RR survey we can derive a linear unbiased estimator for an open survey. Specifically, suppose $e^*(s,z) = b_{s0} + \sum_{i\in s} b_{si}Z_i$ is an unbiased estimator of $T(Y)$ based on an RR survey, i.e.,

$$T(Y) = E_pE_R[b_{s0} + \sum_{i\in s} b_{si}Z_i] = E_p[b^*_{s0} + \sum_{i\in s} b^*_{si}Y_i], \tag{3.3}$$

where $b^*_{s0} = b_{s0} + d_1\sum_{i\in s} b_{si}$ and $b^*_{si} = d_2 b_{si}$. Then, (3.3) shows that $e(s,y) = b^*_{s0} + \sum_{i\in s} b^*_{si}Y_i$ is a linear unbiased estimator of $T(Y)$ based on the corresponding open survey. Thus, for a given sampling design $p(s)$, the two classes of all linear unbiased estimators of π based on an open survey and an RR survey, respectively, are isomorphic.

3.2. Variance Estimation.

We shall now focus on the variance of the estimator $e^*(s,z)$, defined in (3.2). First note that

$$V[Z|Y=1] = \sum_{j=1}^k \alpha_j c_j^2 - (\sum_{j=1}^k \alpha_j c_j)^2 = v_1 \quad \text{and} \quad V[Z|Y=0] = \sum_{j=1}^k \beta_j c_j^2 - (\sum_{j=1}^k \beta_j c_j)^2 = v_0$$

and hence we can write

$$V[Z|Y] = v_0 + (v_1 - v_0)Y. \tag{3.4}$$

Using (3.4), the variance of $e^*(s,z)$ can be written as

$$\begin{aligned} V(e^*(s,z)) &= E_p V_R(e^*(s,z)|s,Y) + V_p E_R(e^*(s,z)|s,Y) \\ &= E_p[\sum_{i\in s} \frac{w_{si}^2}{d_2^2}(v_0 + (v_1 - v_0)Y_i)] + V_p(e(s,y)). \end{aligned} \tag{3.5}$$

The first term in (3.5) is the extra variation due to randomization. Noting that in our application $Y_i^2 = Y_i$, it can be seen that

$$\begin{aligned} V_p(e(s,y)) &= E_p\{e(s,y)\}^2 - \{E_p(e(s,y))\}^2 \\ &= \sum_s w_{s0}^2 p(s) + \sum_{i=1}^N Y_i \sum_{s\ni i}(w_{si}^2 + 2w_{s0}w_{si})p(s) + \sum_{\substack{i,j=1 \\ i\neq j}}^N Y_i Y_j \sum_{s\ni i,j} w_{si}w_{sj}p(s) - (\sum_{i=1}^N Y_i)^2 \\ &= g_0 + \sum_{i=1}^N g_i Y_i + \sum_{\substack{i,j=1 \\ i\neq j}}^N g_{ij} Y_i Y_j, \quad \text{say,} \end{aligned}$$

where,

$$g_0 = \sum_s w_{s0}^2 p(s), \quad g_i = \sum_{s\ni i}(w_{si}^2 + 2w_{s0}w_{si})p(s) - 1, \quad \text{and} \quad g_{ij} = \sum_{s\ni i,j} w_{si}w_{sj}p(s) - 1.$$

Now we discuss how an unbiased estimator of $V(e^*(s,z)$ can be obtained from an unbiased estimator of $V_p(e(s,y)$ based on an open survey. Let d_{si} and d_{sij} be such that

$$\sum_{s\ni i} d_{si}p(s) = g_i, \quad \text{and} \quad \sum_{s\ni i,j} d_{sij}p(s) = g_{ij}$$

so that

$$t(s,y) = g_0 + \sum_{i\in s} d_{si}Y_i + \sum_{\substack{i,j\in s \\ i\neq j}} d_{sij}Y_iY_j$$

is an unbiased estimator of $V_p(e(s,y))$ based on the open survey data. A specific unbiased estimator of $V_p(e(s,y))$ is obtained by using $d_{si} = g_i/\pi_i$ and $d_{sij} = g_{ij}/\pi_{ij}$, where $\pi_i = \sum_{s\ni i} p(s)$

and $\pi_{ij} = \sum_{s \ni i,j} p(s)$. The following result can now be established using the fact that $E_R(U|Y) = Y$.

Theorem 2. An unbiased estimator of $V(e^*(s,z))$, based on RR survey data, is given by

$$t^*(s,z) = g_0 + \sum_{i \in s} d_{si} U_i + \sum_{\substack{i,j \in s \\ i \neq j}} d_{sij} U_i U_j + \sum_{i \in s} \frac{w_{si}^2}{d_2^2}(v_0 + (v_1 - v_0)U_i).$$

Remark 5. So far we have assumed that the randomization probabilities $\{\alpha_j, \beta_j\}$ are the same for all population units. However, in some situations, especially in stratified sampling, as discussed in Kim and Warde (2004) and Christofides (2005), the randomization probabilities may vary over the populations units. We note that the above discussed technique for deriving unbiased estimators based on a RR survey from unbiased estimators based on an open survey also works for varying randomization probabilities. Suppose the randomization probabilities for unit i are $(\vec{\alpha}_i, \vec{\beta}_i) = (\alpha_{i1}, \cdots, \alpha_{ik}, \beta_{i1}, \cdots, \beta_{ik}), i = 1, \cdots, N$. Then letting $d_{i1} = \sum_{j=1}^k \beta_{ij} c_j$, $d_{i2} = \sum_{j=1}^k (\alpha_{ij} - \beta_{ij}) c_j$ and $U_i = (Z_i - d_{i1})/d_{i2}$, it can be seen that $e^*(s,z) = w_{s0} + \sum_{i \in s} w_{si} U_i$ is an unbiased estimator of $T(Y)$. Furthermore, letting

$$v_{i0} = V[Z_i|Y_i = 0] = \sum_{j=1}^k \beta_{ij} c_j^2 - (\sum_{j=1}^k \beta_{ij} c_j)^2 \quad \text{and} \quad v_{i1} = V[Z_i|Y_i = 1] = \sum_{j=1}^k \alpha_{ij} c_j^2 - (\sum_{j=1}^k \alpha_{ij} c_j)^2,$$

we can verify that the variance of $e^*(s,z)$ is

$$V(e^*(s,z)) = E_p[\sum_{i \in s} \frac{w_{si}^2}{d_{i2}^2}(v_{i0} + (v_{i1} - v_{i0})Y_i)] + V_p(e(s,y)). \tag{3.6}$$

and an unbiased estimator of (3.6) is

$$t^*(s,z) = g_0 + \sum_{i \in s} d_{si} U_i + \sum_{\substack{i,j \in s \\ i \neq j}} d_{sij} U_i U_j + \sum_{i \in s} \frac{w_{si}^2}{d_{i2}^2}(v_{i0} + (v_{i1} - v_{i0})U_i).$$

We may also mention that following Chaudhuri (2001, 2004), one can express $V(e^*(s,z))$ in other forms and thence obtain other unbiased estimators of it.

3.3. An Arbitrariness of $e^*(s,z)$.

We note that the construction of $e^*(s,z)$ from a given open survey estimator $e(s,y)$, discussed above, depends on the numerical values $c_1,\cdots,c_k$ used to label the response categories of the RR procedure. For a given sampling design $p(s)$ and a given unbiased estimator for the open survey, the technique yields many unbiased estimators for an RR survey (with $k \geq 3$), by associating different sets of numbers to the response categories. To put this in another way, let $c_i^* = \psi(c_i), i = 1,\cdots,k$ be a transformation of $\{c_1,\cdots,c_k\}$. Then it can be seen that $E[\psi(Z)|Y] = d_1^* + d_2^* Y$, where $d_1^* = \sum_{j=1}^k \beta_j \psi(c_j)$ and $d_2^* = \sum_{j=1}^k (\alpha_j - \beta_j)\psi(c_j)$ and if $d_2^* \neq 0$, letting $U^* = (\psi(Z) - d_1^*)/d_2^*$, it follows that $E_R(U^*|Y) = Y$, and

$$e^{**}(s,z) = w_{s0} + \sum_{i\in s} w_{si} U_i^* = w_{s0}^{**} + \sum_{i\in s} w_{si}^{**} \psi(Z_i),$$

where $w_{s0}^{**} = w_{s0} - (d_1^*/d_2^*)\sum_{i\in s} w_{si}$ and $w_{si}^{**} = w_{si}/d_2^*$, is an unbiased estimator of $T(Y)$ based on the RR survey. Thus, from a given $e(s,y)$, we can construct many unbiased estimators of $T(Y)$ based on the RR survey, by employing different transformations $\psi(.)$.

In view of the above discussion, we may choose $c_1,\cdots,c_k$ to minimize the variance in (3.5). Since the second term of (3.5) does not depend on $c_1,\cdots,c_k$ we need to consider only the first term. First, it is seen easily that $(Z - d_1)/d_2$ and $e^*(s,y)$ are invariant under location and scale transformations of $c_1,\cdots,c_k$, i.e., under $c_i \to \gamma c_i + \delta, i = 1,\cdots,k$, for all γ, δ. From this, it can also be seen that for $k = 2$, i.e., for a $(2 \to 2)$ RR design, $e^*(s,z)$ is unique, independent of the choice of c_1 and c_2. So the estimation methods discussed earlier is well defined for $(2 \to 2)$ RR designs. For $k \geq 3$, without loss of generality (in view of the above mentioned invariance under location and scale transformations), we impose the restrictions:

$$\sum_{i=1}^k \alpha_i c_i = 1 \quad \text{and} \quad \sum_{i=1}^k \beta_i c_i = 0. \tag{3.7}$$

Then, $d_1 = 0, d_2 = 1, v_0 = \sum_{i=1}^{k} c_i^2 \beta_i$ and $v_1 = \sum_{i=1}^{k} c_i^2 \alpha_i - 1$, and the first term of (3.5) reduces to

$$\begin{aligned} E_p[\sum_{i \in s} \frac{w_{si}^2}{d_2^2}(v_0 + (v_1 - v_0)Y_i)] &= v_0 \sum_s \{\sum_{i \in s} w_{si}^2\} p(s) + (v_1 - v_0) \sum_{i=1}^{N} Y_i \sum_{s \ni i} w_{si}^2 p(s) \\ &= A_1(\sum_{i=1}^{k} c_i^2 \alpha_i - 1) + A_2(\sum_{i=1}^{k} c_i^2 \beta_i), \end{aligned} \tag{3.8}$$

where

$$A_1 = \sum_{i=1}^{N} Y_i \sum_{s \ni i} w_{si}^2 p(s) \quad \text{and} \quad A_2 = \sum_s \{\sum_{i \in s} w_{si}^2\} p(s) - A_1.$$

Now, using Lagrangian multipliers, it can be seen that (3.8) is minimized, subject to (3.7), by

$$c_i = \frac{D_2 \alpha_i - D_3 \beta_i}{(D_1 D_2 - D_3^2)(A_1 \alpha_i + A_2 \beta_i)}, \quad i = 1, \cdots, k, \tag{3.9}$$

where,

$$D_1 = \sum_{i=1}^{k} \frac{\alpha_i^2}{A_1 \alpha_i + A_2 \beta_i}, \quad D_2 = \sum_{i=1}^{k} \frac{\beta_i^2}{A_1 \alpha_i + A_2 \beta_i} \quad \text{and} \quad D_3 = \sum_{i=1}^{k} \frac{\alpha_i \beta_i}{A_1 \alpha_i + A_2 \beta_i}.$$

The optimum values in (3.9) depend on the sampling design $p(s)$, the estimator $e(s, y)$ for the open survey and also on $Y_1, \cdots, Y_N$, which are unknown. However, the $\{c_i\}$ in (3.9) depend on $Y_1, \cdots, Y_N$ only through A_1, which may be approximated by

$$A_1 \approx \pi \sum_{i=1}^{N} \sum_{s \ni i} w_{si}^2 p(s).$$

4. Comparison of RR Procedures.

For comparing two RR procedures one should examine both statistical information and respondents' privacy offered by the two procedures. Several authors, including Leysieffer and Warner (1976), Lanke (1976) and Fligner et al. (1977), suggested to compare statistical efficiency, e.g.,

variances of the estimators, of competing procedures while requiring them to offer the same degree of respondents' protection. We shall adopt this approach. The variance of an estimator depends not only on the RR design but also on the choice of the estimator. So for comparing statistical efficiency of two designs, it may be more appropriate to compare some measure of "statistical information" afforded by the two designs. In the following, we shall use Fisher information to compare statistical efficiencies of two RR designs.

We now suggest a criterion for controlling respondents' protection. We start our deliberation with the posterior probabilities in (2.1), which are the determinants of respondents' privacy. The response c_j alters the probability of the respondent's belonging to A by the factor $r_j = P(A|Z = c_j)/\pi$. The ratio r_j may be taken as a measure of respondents' hazard yielded from reporting the response c_j. Clearly, the respondents' hazards $r_1, \cdots, r_k$ corresponding to the responses $c_1, \cdots, c_k$ may be different and a response c_j is hazardous only if $r_j > 1$. Logically, a $(2 \to k)$ RR design is totally non-hazardous if all posterior probabilities equal the prior probability (π), i.e., $r_1 = \cdots = r_k = 1$. However, it can be seen that $r_1 = \cdots = r_k = 1$ if and only if $\alpha_j = \beta_j$ for $j = 1, \cdots, k$, in which case, $\theta_j = P(Z = c_j), j = 1, \cdots, k$, are independent of π and hence the data do not contain any information on π. Thus, to be statistically useful, the design cannot be totally non-hazardous and some r_j must be greater than one. Let $(\alpha_{i1}, \cdots, \alpha_{ik}, \beta_{i1}, \cdots, \beta_{ik}), i = 1, 2$ be two $(2 \to k)$ RR designs with respondents' hazards $(r_{i1}, \cdots, r_{ik}), i = 1, 2$. Strictly speaking, these two RR designs offer equal respondents' protection if and only if $r_{1j} = r_{2j}, j = 1, \cdots, k$. However, this can be satisfied if and only if $\alpha_{1j} = \alpha_{2j}$ and $\beta_{1j} = \beta_{2j}$ for $j = 1, \cdots, k$ (assuming that the design parameters are specified uniquely following a convention, as discussed in Remark 4), i.e., the two designs are the same. Thus, to proceed further we need to employ a weaker criterion for defining equal respondents' protection.

It seems sensible to work with the maximum respondents' hazard: $MRH = \max\{P(A|Z = c_1)/\pi, \cdots, P(A|Z = c_k)/\pi\} = \max\{r_1, \cdots, r_k\}$. This MRH measure is similar to the "primary protection" measure of Fligner et al. (1977). Two RR designs will be considered to offer equal respondents' protection if they have the same MRH value. For controlling respondents' protection, it seems sensible to require the MRH value to be less than a pre-specified number. Several authors, e.g., Anderson (1976), Lanke (1976), Leysieffer and Warner (1976) and Fligner et al. (1977), essentially suggested this approach. However, the values of r_j and hence MRH depend on the unknown parameter π. Thus, in practice, we would need to put an upper bound on MRH for a specific value of π. Since $P(A|Z = c_j)/\pi$ is an increasing function of (α_j/β_j), putting an upper bound on MRH, for a specific value of π, is equivalent to putting an upper bound on $\max\{\alpha_1/\beta_1, \cdots, \alpha_k/\beta_k\}$. Thus, we shall take

$$R(\vec{\alpha}, \vec{\beta}) = \max\{\frac{\alpha_1}{\beta_1}, \cdots, \frac{\alpha_k}{\beta_k}\} \tag{4.1}$$

as our measure of the degree of privacy afforded by the $(2 \to k)$ RR design $(\vec{\alpha}, \vec{\beta})$. It can be seen that two RR designs $(\vec{\alpha}_1, \vec{\beta}_1)$ and $(\vec{\alpha}_2, \vec{\beta}_2)$ have a common value of MRH if and only if they have same value for R, i.e., $R(\vec{\alpha}_1, \vec{\beta}_1) = R(\vec{\alpha}_2, \vec{\beta}_2)$. Noting that

$$\frac{\alpha_j}{\beta_j} = \frac{P(Z = c_j|A)}{P(Z = c_j|A^c)} = \frac{P(A|Z = c_j)/P(A)}{P(A^c|Z = c_j)/P(A^c)}, \quad j = 1, \cdots, k,$$

the ratios $\{\alpha_j/\beta_j\}$ may be regarded as Bayes factors. They were used by Leysieffer and Warner (1976) in their discussion of respondents' protection.

In summary, for comparing two procedures, we suggest to hold the privacy measure in (4.1) equal for the two procedures and then compare statistical efficiency, measured by Fisher information. Using this approach, we shall next show that for any $(2 \to k)$ design with $k \geq 3$, there exists a better $(2 \to 2)$ design. The result also helps us to identify the most efficient RR

design at any given level of privacy protection.

Theorem 3. Let D be any $(2 \to k)$ RR design with $k \geq 3$ and randomization parameters $(\vec{\alpha}, \vec{\beta})$. Then, there exists a $(2 \to 2)$ RR design D_0 which provides the same respondents' protection, measured by (4.1), as D and at least as much statistical information as D.

Proof. For notational simplicity, without loss of generality suppose that $\alpha_1/\beta_1 = \max\{\alpha_j/\beta_j\}$. If $(\alpha_1/\beta_1) = 1$, the data do not contain any information on π. So, we shall only consider the case of $(\alpha_1/\beta_1) > 1$. Let $b_0 = \beta_1/\alpha_1 (< 1)$ and D_0 be the $(2 \to 2)$ RR design with response variable V and $P(V = 1|A) = 1$ and $P(V = 1|A^c) = b_0$. It is easy to verify that D_0 and D offer the same degree of respondents' protection, as measured by (4.1).

Next we shall show that D is equivalent to post-randomizing the data generated by D_0. Let $\vec{\gamma} = (\gamma_1, \cdots, \gamma_k) = (\vec{\beta} - b_0\vec{\alpha})/(1 - b_0)$, and randomly transform V to $c_1, \cdots, c_k$ according to the probabilities $P(c_j|V = 1) = \alpha_j$ and $P(c_j|V = 0) = \gamma_j$ for $j = 1, \cdots, k$, and denote the resulting variable by Z. From the fact that $\alpha_1/\beta_1 \geq \alpha_i/\beta_i, i = 1, \cdots, k$, it can be checked easily that $\gamma_i \geq 0, i = 1, \cdots, k$ and $\sum_{i=1}^{k} \gamma_i = 1$, i.e., $\vec{\gamma}$ is a probability vector. We may also note that the transformation of V to Z is performed without using the true category of the respondent or the true value of π. Now, it follows easily that for $i = 1, \cdots, k$,

$$P(Z = c_i|A) = P(Z = c_i|V = 1)P(V = 1|A) + P(Z = c_i|V = 0)P(V = 0|A) = \alpha_i \qquad (4.2)$$

and

$$P(Z = c_i|A^c) = P(Z = c_i|V = 1)P(V = 1|A^c) + P(Z = c_i|V = 0)P(V = 0|A^c) = \beta_i. \qquad (4.3)$$

So, generating data using D is equivalent to first generating data using D_0 and then randomizing them using the known probabilities $\{\alpha_i\}$ and $\{\gamma_i\}$.

In effect, D adds "random noise" to the data generated by D_0, from which it is quite intuitive that D_0 is more informative than D. This can be established formally following Anderson (1977), as discussed below. Let $I_V(\pi), I_Z(\pi)$ and $I_{VZ}(\pi)$ denote Fisher's information on π contained in V, Z and (V, Z), respectively. Then from general properties of Fisher's information it follows that

$$I_{VZ}(\pi) = I_V(\pi) + I_{Z|V}(\pi) = I_Z(\pi) + I_{V|Z}(\pi), \tag{4.4}$$

where $I_{Z|V}(\pi)$ is the average conditional information in Z given V and $I_{V|Z}(\pi)$ is defined similarly. Since the conditional distribution of Z given V does not depend on π, $I_{Z|V}(\pi) = 0$ and (4.4) implies that

$$I_V(\pi) - I_Z(\pi) = I_{V|Z}(\pi) \geq 0,$$

which completes the proof of the theorem.

The proof of Theorem 3 shows that D is as informative as D_0 only when $I_{V|Z}(\pi) = 0$, or equivalently, the conditional distribution of V given Z is independent of π. Note that

$$P_\pi(V = 1|Z = c_j) = \frac{\alpha_j\pi + \alpha_j(\beta_1/\alpha_1)(1-\pi)}{\alpha_j\pi + \beta_j(1-\pi)}$$

is independent of π if and only if $\alpha_j(\beta_1/\alpha_1) = \beta_j$, i.e., $\alpha_j/\beta_j = \alpha_1/\beta_1$. So, the conditional distribution of V given Z is independent of π if and only if $\alpha_1/\beta_1 = \cdots = \alpha_k/\beta_k$, i.e., $\alpha_j = \beta_j$ for $j = 1, \cdots, k$, in which case D is non-informative. Thus, for any informative design D, $I_{V|Z}(\pi) > 0$ and hence D_0 is more informative than D.

Our proof of Theorem 3 is constructive; we not only show existence of a better design D_0 but also provide a recipe for finding one. The main implication of Theorem 3 is that for surveying dichotomous populations one should use only binary response variables, i.e., use only $(2 \to 2)$ RR designs. Then, Nayak's (1994) admissibility result (see, Remark 2) suggests that one should

use only $(2 \to 2)$ RR designs with $P(Yes|A) = 1$. The design D_0, constructed in the proof of Theorem 3, is an admissible design. From Nayak (1994) and our Theorem 3 it can be seen that the best RR design at a specified level ($r \geq 1$) of the privacy measure $R(\vec{\alpha}, \vec{\beta})$ is the $(2 \to 2)$ design with $\alpha_1 = 1, \alpha_2 = 0, \beta_1 = 1/r$ and $\beta_2 = 1 - 1/r$.

Remark 6. Information domination of D_0 over D can also be seen using Blackwell's (1951) ideas for comparing statistical experiments. Equations (4.2) and (4.3) show that D_0 is sufficient for D, by Blackwell's definition of sufficiency. Then, from Blackwell (1953) it follows that for every loss function $L(\pi, \hat{\pi})$ and any estimator $\hat{\pi}_D$ based on D, there exists an estimator $\hat{\pi}_*$ based on D_0 such that $E[L(\pi, \hat{\pi}_*)] \leq E[L(\pi, \hat{\pi})]$ for all $0 \leq \pi \leq 1$.

Remark 7. We may note that some papers, e.g., Greenberg et al. (1969) and Mangat and Singh (1990), compared variances without holding respondents' protection equal. Similarly, Christofides (2003) compared the variances of Warner's estimator and his estimator, based on his $(2 \to k)$ procedure, not taking respondents' protection into account. In his design, $\beta_i = \alpha_{(k-i+1)}, i = 1, \cdots, k$, and he showed that for any Warner's design with given p, one can find suitable values of the parameters $\{\alpha_i\}$ of his design, with $k \geq 3$, such that the variance of his estimator is less than that of Warner's estimator. Thence he concluded that his RR technique "improves upon" Warner's procedure. As an illustrative example, he took the Warner's estimator with $p = 0.6$, in which case

$$Var(\hat{\pi}_W) = \frac{\pi(1-\pi)}{n} + \frac{6}{n}. \tag{4.5}$$

Then he showed that his estimators, with $k = 6$ and $(\alpha_1, \cdots, \alpha_6) = (0.38, 0.02, 0.19, 0.1, , 0.05, 0.26)$, has variance

$$Var(\hat{\pi}_C) = \frac{\pi(1-\pi)}{n} + \frac{3.76}{n},$$

which is clearly less than $Var(\hat{\pi}_W)$ in (4.5) for all $0 \leq \pi \leq 1$. However, we find Christofides' (2003) argument and his conclusion to be flawed. First, he did not take respondents' protection into account. Second, if only the variances are compared, the fact that for any Warner's estimator, with given p, *there exists* a Christofides' estimator with smaller variance does not validate the conclusion that Christofides' *procedure* is better than Warner's *procedure*. This is because it can also be seen that for any given Christofides' estimator, *there exists* a Warner's estimator, with suitable choice of p, with uniformly smaller variance; note that the last term of (1.1) can be made arbitrarily small because it approaches 0 as p tends to 1 (or 0). For example, the Warner's estimator with $p = .65$, in which case the last term of (1.1) is $2.528/n < 3.76/n$, is better (in terms of variance) than the Christofides' estimator considered in the illustrative example.

5. Discussion.

In this paper we presented a unified framework for discussing RR surveys of dichotomous populations with multiple response categories. Several RR procedures have appeared in the literature, but typically, each procedure has been discussed within its own framework. We believe, this has hindered systematic thinking about the core statistical issues and has led to erroneous conclusions. A common framework is helpful, and perhaps necessary, for abstraction and formalization of the key elements relating to respondents' privacy and statistical efficiency and comparison of various procedures. We hope the ideas put forward in this paper will be helpful in developing a unified theory of RR surveys, comparing various procedures and reaching valid conclusions. While we discussed only polychotomous response variables, we believe our ideas can be extended to RR surveys of dichotomous populations with quantitative response variables, such as

the procedures discussed in Franklin (1989) and Chua and Tsui (2000).

A unified theory is also helpful for recognizing connections of RR surveys to other areas of statistics, such as comparison of experiments (see Remark 6) and estimation from open surveys. RR surveys are closely related to the post-randomization method (PRAM) for controlling statistical disclosure. The PRAM, introduced by Gouweleeuw et al. (1998), is concerned with protecting respondents' privacy while releasing microdata (already collected) for public use. It stochastically transforms the values of categorical variables in a data set using a known Markov matrix. As noted by Van den Hout and Van der Heijden (2002), mathematically, the PRAM is equivalent to an RR procedure. Both are concerned with protection of respondents' privacy and statistical efficiency; only difference is that in RR surveys, the responder randomizes the response at data gathering stage whereas in PRAM randomization is carried out by the surveyor after the data are collected. Thus, the results for RR surveys can be used beneficially in statistical disclosure control.

Acknowledgment. The research of Tapan Nayak was supported in part by a grant from the Institute for Integrating Statistics in Decision Sciences of The George Washington University.

References

[1] Anderson, H. (1976). Estimation of a proportion through randomized response. *Int. Stat. Rev.*, 44, 213-217.

[2] Anderson, H. (1977). Efficiency versus protection in a general randomized response model. *Scand. J. Statist.*, 4, 11-19.

[3] Blackwell, D. (1951). Comparison of experiments, *Proceedings of Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, pp. 93-102.

[4] Blackwell, D. (1953). Equivalent comparison of experiments, *Ann. Math. Statist.*, 24, 265-272.

[5] Chaudhuri, A. (2001). Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population. *J. Statist. Plann. Inference*, 94, 37-42.

[6] Chaudhuri, A. (2004). Christofides' randomized response technique in complex sample surveys. *Metrika*, 60, 223-228.

[7] Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Response: Theory and Techniques*. Marcel Dekker, New York.

[8] Christofides, T.C. (2003). A generalized randomized response technique. *Metrika*, 57, 195-200.

[9] Christofides, T.C. (2005). Randomized response in stratified sampling. *J. Statist. Plann. Inference*, 128, 303-310.

[10] Chua, T.C. and Tsui, A.K. (2000). Procuring honest responses indirectly. *J. Statist. Plann. Inference*, 90, 107-116.

[11] Fligner, M.A., Policello, G.E. and Singh, J. (1977). A comparison of two randomized response survey methods with consideration for the level of respondent protection. *Commun. Statist. - Theory Meth,*, 6, 1511-1524.

[12] Franklin, L.A. (1989). Randomized response sampling from dichotomous populations with continuous randomization. *Survey Methodology*, 15, 225-235.

[13] Greenberg, B.G., Abul-Ela, A-L. A., Simmons, W.R. and Horvitz, D.G. (1969). The unrelated question randomized response model: theoretical framework. *J. Amer. Statist Assoc.*, 64, 520-539.

[14] Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J. and De Wolf, P.P. (1998). Post randomisation for statistical disclosure control: Theory and implementation. *J. Official Statist.*, 14, 463–478.

[15] Kim, J-M. and Warde, W.D. (2004). A stratified Warner's randomized response model. *J. Statist. Plann. Inference*, 120, 155-165.

[16] Kuk, A.Y.C. (1990). Asking sensitive questions indirectly. *Biometrika*, 77, 436-438.

[17] Lanke, J. (1976). On the degree of protection in randomized interviews. *Int. Stat. Rev.*, 44, 197-203.

[18] Leysieffer, R.W. and Warner, S.L. (1976). Respondent jeopardy and optimal designs in randomized response models. *J. Amer. Statist Assoc.*, 71, 649-656.

[19] Mangat, N.S. (1994). An improved randomized response strategy. *J. R. Statist. Soc.*, 56, 93-95.

[20] Mangat, N.S. and Singh, R. (1990). An alternative randomized response procedure. *Biometrika*, 77, 439-442.

[21] Nayak, T.K. (1994). On randomized response surveys for estimating a proportion. *Commun. Statist. - Theory Meth,*, 23, 3303-3321.

[22] Padmawar, V.R. and Vijayan, K. (2000). Randomized response revisited. *J. Statist. Plann. Inference*, 90, 293-304.

[23] Van den Hout, A. Van der Heijden, P.G.M. (2002). Randomized response, statistical disclosure control and misclassification: a review. *Int. Stat. Rev.*, 70, 269-288.

[24] Warner, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *J. Amer. Statist Assoc.*, 60, 63-69.