Multi-stage Optional Unrelated Question RRT Model

Anu Chhabra

Department of Mathematics, Lakshmibai College, University of Delhi, India
a.chhabra02@gmail.com

B. K. Dass

Department of Mathematics, University of Delhi, India
dassbk@rediffmail.com

Samridhi Mehta

Department of Mathematics, Hindu College, University of Delhi, India
mehta.samridhi@gmail.com

Received 28 September 2014

Accepted 23 September 2015

Abstract

Sihm et al. (2014) introduced modified optional unrelated question RRT model in both binary and quantitative response situations wherein the prevalence of the sensitive variable and the sensitivity level of the underlying sensitive question could be estimated simultaneously without using a split sample approach. In this study, we propose a three−stage optional unrelated question RRT model for both binary and quantitative response situations which combines the essence of Sihm et al. (2014) model and the three−stage optional additive RRT model proposed by Mehta et al. (2012). The efficiencies of Sihm et al. (2014) model and proposed three−stage optional unrelated question RRT model are compared using simulations. The privacy measures of the two models in question are also discussed. Comparisons for the binary models are based on Lanke (1976) measure while Yan et al. (2009) measure is used to compare the privacy measures for the quantitative models.

Keywords: Unrelated Question RRT Models, Optional RRT Models, Simulation Study, Parameter Estimation and Privacy Protection.

2000 Mathematics Subject Classification: 62D05

1. Introduction

It has been often observed in interviews that people tend to evade those questions in face-to-face surveys whose truthful answers have social desirability implications and/or can attract punitive action. Such questions could be on sensitive subjects such as sexual behaviour, drug use, abortions, spousal abuse, and many others. Innocuous questions ordinarily receive honest responses, but questions requiring personal or controversial traits induce resistance in the form of refusal to answer the question or reporting of an untruthful response. Warner (1965) suggested a randomized response model to overcome this problem of social desirability bias. It is a research method which allows the respondents to respond to sensitive issues while maintaining confidentiality. Greenberg et al. (1969) introduced the unrelated question model for binary response situations. Instead of requesting the respondent to reply affirmatively or negatively to the sensitive research

Published by Atlantis Press

question, the alternatives are stated differently in this model. Here, the respondent face the randomization device in which the sensitive question is asked with known probability p and an innocuous question which has no possible embarrassment is asked with probability $(1-p)$. Gupta et al. (2002) introduced the concept of optional models in RRT wherein an option was given to the respondents. They can either answer the research question truthfully if they do not find it sensitive or they can provide a scrambled response using a randomization device if they feel the research question is sensitive. Several modifications of the Gupta et al. (2002) model exist in literature.

Sihm et al. (2014) proposed modified optional unrelated question models − a binary and a quantitative RRT model, which offer the respondents the option of answering the sensitive question directly if they find the research question non−sensitive. In an optional model of this type, there are two parameters of interest −the sensitivity level of the question (i.e. the proportion of respondents in the population who consider the question sensitive), and prevalence of the sensitive characteristic in the population. The prevalence of the sensitive characteristic is estimated by using an optional unrelated question RRT model but the sensitivity level of the question is estimated from the sample by using the traditional Greenberg et al. (1969) model. This eliminates the need for a split sample approach which requires a larger total sample size.

In our current work, we propose a three−stage optional unrelated question RRT model that does not involve the split sample approach. In the proposed study, the respondents are asked two questions. A randomly selected proportion T of respondents answer the main research question truthfully and a known proportion F of respondents provide a randomized response to the research question using Greenberg et al. (1969) model in binary response situations (or Greenberg et al. (1971) model in quantitative response situations) in which the respondent uses the randomization device which bears the sensitive question with known probability p and an unrelated innocuous question which has no possible embarrassment, is answered with probability $(1-p)$. The remaining $(1-T-F)$ proportion of respondents use Gupta et al. (2013) optional unrelated question RRT model, in which the respondents are given the option to answer the research question directly if they don't find the research question sensitive. In case the respondents feel that the research question is sensitive, then they answer it using the Greenberg et al. (1969) model or Greenberg et al. (1971) model (depending upon the type of response). The split sample approach which requires a larger total sample size is not used. Instead, we estimate the sensitivity level from the same sample by using the Greenberg et al. (1969) model. In this paper, both the binary response and the quantitative response situations are dealt with and estimates for the prevalence of sensitive behaviour π in binary response situation and the mean response μ_X of the quantitative sensitive question are obtained.

In Section 2, the theoretical framework for the three − stage binary model and quantitative models is discussed. The simulation study is discussed in Section 3 and it helps validate our findings of Section 2. The aspect of privacy protection of respondents is discussed in Section 4. The binary response models are compared using the Lanke (1976) measure and the quantitative response models are compared using the Yan et al. (2009) measure. The results are summarized in Section 5.

2. Proposed Three-Stage Models

In this section, we propose two optional RRT models‑ one for binary response situations and one for quantitative response situations. The proposed models are optional three−stage RRT models and are an extension of two-stage model proposed by Mangat and Singh (1990) but in the context of Unrelated Question models.

2.1 *Three-Stage Binary Model*

In the proposed binary model, all the respondents are asked two questions. The question about sensitivity is asked first via randomization device 1. In this randomization, the sensitive question is "Is the main research question sensitive?" It is asked along with an unrelated innocuous question. The underlying sensitivity level w and its variance can be estimated from the sample by using the Greenberg et al. (1969) model. Then, in the model, all the respondents of the same sample are asked another question which is to ascertain the prevalence of the sensitive characteristic in the population using randomization device 2.

In randomization device 2, the same sample is used. A known proportion T of respondents answer the research question truthfully and a known proportion F of respondents provide a randomized response using the Greenberg et al. (1969) model in which the respondent uses the randomization device which bears sensitive question with the known probability p_b and an unrelated innocuous question which has no possible embarrassment, is answered with probability $1 - p_b$. The remaining proportion $(1 - T - F)$ of respondents uses Gupta et al. (2013) optional unrelated question model, in which the respondent is given the option to answer the research question directly (or using the Greenberg et al. (1969) model with known parameter p_b) if they find the research question non−sensitive (or sensitive).

We use the following notation:

- n be the sample size,
- π_a be the known probability of an unrelated question used in Greenberg et al. (1969) model of randomization device 1,
- π_b be the known probability of another unrelated question used in Greenberg et al. (1969) model of randomization device 2,
- π be the unknown proportion of population that belongs to the sensitive group,
- p_a be the known probability of the respondent selecting the question about sensitivity in Greenberg et al. (1969) model of randomization device 1,
- p_b be the known probability of the respondent selecting the sensitive question in Greenberg et al. (1969) model of randomization device 2,
- w be the sensitivity level of the main research question in the population,
- P_{y1} be the probability of 'yes' response from a respondent in randomization device 1.
- P_{y2} be the probability of 'yes' response from a respondent in randomization device 2.

Thus, using randomization device 1 we obtain,

$$P_{y1} = p_a w + (1 - p_a)\pi_a \tag{1}$$

Solving for w,

$$w = \frac{P_{y1} - (1 - p_a)\pi_a}{p_a} \tag{2}$$

Thus, the estimator of w is given by,

$$\hat{w} = \frac{\hat{P}_{y1} - (1 - p_a)\pi_a}{p_a} \tag{3}$$

where $\hat{P}_{y1}$ is the proportion of 'yes' responses in randomization device 1.

It is clear that $\hat{w}$ is an unbiased estimator of w with

$$\text{var}(\hat{w}) = \frac{P_{y1}(1 - P_{y1})}{np_a^{\ 2}} \tag{4}$$

We also have,

$$P_{y2} = T\pi + F\{p_b\pi + (1-p_b)\pi_b\} + (1-T-F)\{(1-w)\pi + w[p_b\pi + (1-p_b)\pi_b]\} \tag{5}$$

Equation (5) can be rearranged as,

$$P_{y2} = \{(1-w)[1-F(1-p_b)] + w[p_b + T(1-p_b)]\}\pi + (1-p_b)\pi_b\{F + w(1-T-F)\} \tag{6}$$

Solving for π,

$$\pi = \frac{P_{y2} - (1-p_b)\pi_b\{F + w(1-T-F)\}}{(1-w)\{1-F(1-p_b)\} + w[T(1-p_b) + p_b]} \tag{7}$$

Thus we have an estimator of π given by,

$$\hat{\pi} = \frac{\hat{P}_{y2} - (1-p_b)\{F + \hat{w}(1-T-F)\}\pi_b}{(1-\hat{w})\{1-F(1-p_b)\} + \hat{w}[T(1-p_b) + p_b]} \tag{8}$$

where $\hat{w}$ is obtained from equation (3) and $\hat{P}_{y2}$ is the proportion of 'yes' responses from randomization device 2.

After applying first order Taylor's expansion to equation (8), we have

$$\hat{\pi} = \hat{\pi}(P_{y2}, w) + \left.\frac{\partial\hat{\pi}(\hat{P}_{y2}, \hat{w})}{\partial\hat{P}_{y2}}\right|_{(P_{y2},w)} .(\hat{P}_{y2} - P_{y2}) + \left.\frac{\partial\hat{\pi}(\hat{P}_{y2}, \hat{w})}{\partial\hat{w}}\right|_{(P_{y2},w)} .(\hat{w} - w)$$

which gives

$$\hat{\pi} = \frac{P_{y2} - (1-p_b)\pi_b\{F + w(1-T-F)\}}{(1-w)\{1-F(1-p_b)\} + w[T(1-p_b) + p_b]} + A(\hat{P}_{y2} - P_{y2}) + B(\hat{w} - w) \tag{9}$$

where

$$A = \left.\frac{\partial\hat{\pi}(\hat{P}_{y2}, \hat{w})}{\partial\hat{P}_{y2}}\right|_{(P_{y2},w)} = \frac{1}{\{(1-w)[1-F(1-p_b)] + w[T(1-p_b) + p_b]\}} \tag{10}$$

and

$$B = \left.\frac{\partial\hat{\pi}(\hat{P}_{y2}, \hat{w})}{\partial\hat{w}}\right|_{(P_{y2},w)} = \frac{[-(1-p_b)\pi_b(1-T-F)][(1-w)\{1-F(1-p_b)\} + w\{T(1-p_b) + p_b\}]}{\{(1-w)[1-F(1-p_b)] + w[T(1-p_b) + p_b]\}^2}$$

$$- \frac{[P_{y2} - (1-p_b)\pi_b\{F + w(1-T-F)\}]\{F(1-p_b) - 1 + p_b + T(1-p_b)\}}{\{(1-w)[1-F(1-p_b)] + w[T(1-p_b) + p_b]\}^2} \tag{11}$$

The expectation of $\hat{\pi}$, to first order of approximation, is given by,

$$E(\hat{\pi}) = \frac{P_{y2} - (1-p_b)\pi_b\{F + w(1-T-F)\}}{(1-w)\{1-F(1-p_b)\} + w[T(1-p_b) + p_b]} = \pi \tag{12}$$

Thus, using (4), we have the following result:

Theorem 1: *Up to first order Taylor's Approximation, $\hat{\pi}$ is an unbiased estimator of π with*

$$Var(\hat{\pi}) = A^2 \frac{P_{y2}(1-P_{y2})}{n} + B^2 \frac{P_{y1}(1-P_{y1})}{np_a^2}$$

where A and B are as given in (10) *and* (11).

In the next section, we propose the quantitative version of the above model.

2.2 *Three-Stage Quantitative Model*

The proposed quantitative version of the model is on the lines of binary model given above and again, all the respondents are asked two questions. The question about sensitivity is asked first using randomization device 1 which is same as the one used in the binary case above. All the respondents of the same sample are asked another question to ascertain the mean prevalence of the sensitive characteristic *X* in the population. This is done using randomization device 2 using the same sample as the one used for the first question. With this randomization device, a known proportion *T* of respondents answer the research question truthfully and a known proportion *F* of respondents provide a randomized response using Greenberg et al. (1971) model in which the respondent uses the randomization device which bears sensitive question with the known probability p_b and an unrelated innocuous question is answered with probability $1 - p_b$. The remaining proportion$(1 - T - F)$of respondents use Gupta et al. (2013) optional unrelated question model, in which the respondent is given the option to answer the research question directly (or by using the Greenberg et al. (1971) model with known parameter p_b, and known mean and variance of the innocuous variable) if they find the research question non−sensitive (sensitive).

We use the following notation:

- w be the sensitivity level of the survey question in the population,
- n be the sample size,
- π_a be the known probability of an unrelated question used in Greenberg et al. (1969) model of randomization device 1,
- p_a be the known probability of the respondent selecting the question about sensitivity in Greenberg et al. (1969) model of randomization device 1,
- p_b be the known probability of the respondent selecting the sensitive question in Greenberg et al. (1971) model of randomization device 2,
- μ_Y and σ_Y^2 be the known mean and variance of an innocuous question used in Greenberg et al. (1971) model,
- μ_X and σ_X^2 be the unknown mean and variance of the sensitive question of the population,
- Z be the reported quantitative response to randomization device 2 by a respondent.

Then from randomization device 1, we obtain $\hat{w}$ as an unbiased estimator of w and $\hat{w}$ and its variance are given in (3) and (4) above respectively.

From randomization device 2, we get

$$Z = \begin{cases} X \\ Y \end{cases} \text{with probability} \begin{cases} T + Fp_b + (1-T-F)\left[(1-w) + wp_b\right] & \text{(sensitive question)} \\ F\left(1-p_b\right) + (1-T-F)w(1-p_b) & \text{(innocuous question)} \end{cases}$$

Thus,

$$E(Z)=E(X)\{T+Fp_b+(1-T-F)(1-w+wp_b)\}\\ +E(Y)[F(1-p_b)+(1-T-F)w(1-p_b)] \tag{13}$$

and

$$Var(Z)=\{T+Fp_b+(1-T-F)(1-w+wp_b)\}E(X^2)+\\ [F(1-p_b)+(1-T-F)w(1-p_b)]E(Y^2)-\mu_Z^2 \tag{14}$$

i.e.,

$$Var(Z)=\{T+Fp_b+(1-T-F)(1-w+wp_b)\}\left(\sigma_X^2+\mu_X^2\right)+\\ [F(1-p_b)+(1-T-F)w(1-p_b)]\left(\sigma_Y^2+\mu_Y^2\right)-\mu_Z^2 \tag{15}$$

From equation (13),

$$\mu_Z=\mu_X\{T+Fp_b+(1-T-F)(1-w+wp_b)\}+\mu_Y[F(1-p_b)+(1-T-F)w(1-p_b)] \tag{16}$$

Solving for μ_X,

$$\mu_X=\frac{\mu_Z-\mu_Y[F(1-p_b)+(1-T-F)w(1-p_b)]}{\{T+Fp_b+(1-T-F)(1-w+wp_b)\}} \tag{17}$$

Thus the estimator for μ_X is given by,

$$\hat{\mu}_X=\frac{\hat{\mu}_Z-\mu_Y\left[F(1-p_b)+(1-T-F)\hat{w}(1-p_b)\right]}{\left\{T+Fp_b+(1-T-F)(1-\hat{w}+\hat{w}p_b)\right\}} \tag{18}$$

where $\hat{w}$ is unbiased estimator for w obtained from equation (3) and $\hat{\mu}_Z$ is the estimate for μ_Z obtained from the sample with $Var(\hat{\mu}_Z)=\dfrac{Var(Z)}{n}$.

After applying first order Taylor's expansion to equation (18), we obtain

$$\hat{\mu}_X=\hat{\mu}_X(\mu_z,w)+\left.\frac{\partial\hat{\mu}_X(\hat{\mu}_z,\hat{w})}{\partial\hat{\mu}_z}\right|_{(\mu_z,w)}.(\hat{\mu}_z-\mu_Z)+\left.\frac{\partial\hat{\mu}_X(\hat{\mu}_z,\hat{w})}{\partial\hat{w}}\right|_{(\mu_z,w)}.(\hat{w}-w)$$

i.e.

$$\hat{\mu}_X=\frac{\mu_Z-\mu_Y[F(1-p_b)+(1-T-F)w(1-p_b)]}{\{T+Fp_b+(1-T-F)(1-w+wp_b)\}}+A(\hat{\mu}_z-\mu_Z)+B(\hat{w}-w) \tag{19}$$

where

$$A=\left.\frac{\partial\hat{\mu}_X(\hat{\mu}_z,\hat{w})}{\partial\hat{\mu}_z}\right|_{(\mu_z,w)}=\frac{1}{T+Fp_b+(1-T-F)(1-w+wp_b)} \tag{20}$$

and

$$B=\left.\frac{\partial \hat{\mu}_X(\hat{\mu}_z,\hat{w})}{\partial \hat{w}}\right|_{(\mu_z,w)}=\frac{\{T+Fp_b+(1-T-F)(1-w+wp_b)\}\{-\mu_Y(1-T-F)(1-p_b)\}}{\{T+Fp_b+(1-T-F)(1-w+wp_b)\}^2}$$

$$+\frac{[\mu_Z-\mu_Y\{F(1-p_b)+(1-T-F)w(1-p_b)\}](1-T-F)(1-p_b)}{\{T+Fp_b+(1-T-F)(1-w+wp_b)\}^2} \qquad (21)$$

Thus,

$$E\left(\hat{\mu}_X\right)=\frac{\mu_Z-\mu_Y[F(1-p_b)+(1-T-F)w(1-p_b)]}{\{T+Fp_b+(1-T-F)(1-w+wp_b)\}}=\mu_X \qquad (22)$$

Using the expression of $\mathrm{var}(\hat{w})$ from (4), we get the following result:

Theorem 2: *Up to first order Taylor's Approximation,* $\hat{\mu}_X$ *is an unbiased estimator of* μ_X *with* $Var(\hat{\mu}_X)=A^2\dfrac{Var\,Z}{n}+B^2\dfrac{P_{y1}(1-P_{y1})}{np_a^2}.$

where A and B are as given in (20) *and* (21).

3. Simulation Study

In this section, the theoretical results obtained in Section 2 for our estimators $\hat{\mu}_X$, $\hat{\pi}$ and $\hat{w}$ are verified empirically. All the simulations were conducted using SAS. For the binary response models, parameters T and F , were allowed to vary while all other variables were fixed. We used number of trials = 10000, $w=0.9$, $\pi=0.3$, $\pi_a=0.1$, $\pi_b=0.7$, $p_a=0.5$, $p_b=0.85$, and $n=1000$. For the quantitative response models, parameters T and F were allowed to vary while all other variables were fixed again and we used the number of trials = 10000, $w=0.9$, $\pi_a=0.1$, $p_a=0.5$, $p_b=0.85$, $\mu_X=2$, $\mu_Y=7$ and $n=1000$. Further, X and Y are assumed to follow Poisson distribution with parameters μ_X and μ_Y respectively. Both the proposed models are valid for those combinations of T and F for which $T+F<1$. Thus, in tables given below, the combinations of T and F for which $T+F\geq 1$ are marked with a dash (−).

3.1 *Simulation of $\hat{\pi}$ and $\hat{w}$ for Binary Three-Stage Model*

The simulation results provide strong support to our earlier finding that $\hat{\pi}$ and $\hat{w}$ are unbiased estimators of π and w respectively. The theoretical and simulated variances of $\hat{w}$ are very close and the theoretical and simulated variances of $\hat{\pi}$ get closer as F increases for all values for T . For example: one may observe from Table 1 that for $T=0.3$ and $F=0.3$, the theoretical value of $Var(\hat{w})=0.001$ and simulated value of $Var(\hat{w})=0.0010117$. Similarly, for $T=0.3$ and $F=0.3$, the theoretical value of $Var(\hat{\pi})=0.000276974$

and simulated value of $Var(\hat{\pi}) = 0.000278646$. The first order Taylor's Approximation was used to calculate the theoretical values for $Var(\hat{\pi})$. Note that since the optional unrelated question binary model of Sihm et al. (2014) is a one-stage model ($T = 0, F = 0$), thus, in Table 1 below that model is discussed only for one choice of (*T, F*) value, namely (0,0). Also, in Table 1, (*) indicates that the corresponding quantity is for the Sihm et al. (2014) binary model.

Table 1 : Simulation Results of Binary Model

Trials = 10000, $w = 0.9$, $\pi = 0.3$, $\pi_a = 0.1$, $\pi_b = 0.7$, $p_a = 0.5$, $p_b = 0.85$, and $n = 1000$.						
F	T					
		0	0.1	0.3	0.5	0.7
0	Var($\hat{w}$) Empirical value	0.0010146	0.0010132	0.0010132	0.0010132	0.0010132
	Var($\hat{w}$) Empirical value(*)	0.000974				
	Var($\hat{w}$) Theoretical value	0.001	0.001	0.001	0.001	0.001
	Var($\hat{w}$) Theoretical value(*)	0.001				
	Mean($\hat{w}$) Empirical value	0.9002198	0.8999242	0.8999242	0.8999242	0.8999242
	Mean($\hat{w}$) Empirical value(*)	0.900596				
	Var($\hat{\pi}$) Empirical value	0.000299074	0.000300267	0.000273351	0.000252186	0.000236159
	Var($\hat{\pi}$) Empirical value(*)	0.000303				
	Var($\hat{\pi}$) Theoretical value	0.000310447	0.000298012	0.000274969	0.000254119	0.000235208
	Var($\hat{\pi}$) Theoretical value(*)	0.00031				
	Mean($\hat{\pi}$) Empirical value	0.2997113	0.2999437	0.29992	0.2999305	0.2999561
	Mean($\hat{\pi}$) Empirical value(*)	0.300231				
0.1	Var($\hat{w}$) Empirical value	0.0010132	0.0010117	0.0010117	0.0010117	0.0010117
	Var($\hat{w}$) Theoretical value	0.001	0.001	0.001	0.001	0.001
	Mean($\hat{w}$) Empirical value	0.8999242	0.9001588	0.9001588	0.9001588	0.9001588
	Var($\hat{\pi}$) Empirical value	0.000314227	0.000298961	0.000276951	0.000256815	0.000237076
	Var($\hat{\pi}$) Theoretical value	0.000310843	0.000298472	0.000275547	0.000254804	0.000235989
	Mean($\hat{\pi}$) Empirical value	0.2999546	0.2998695	0.2998338	0.2997007	0.2994791

0.3	Var($\hat{w}$) Empirical value	0.0010132	0.0010117	0.0010117	0.0010117	–
	Var($\hat{w}$) Theoretical value	0.001	0.001	0.001	0.001	–
	Mean($\hat{w}$) Empirical value	0.8999242	0.9001588	0.9001588	0.9001588	–
	Var($\hat{\pi}$)Empirical value	0.000315784	0.000300742	0.000278646	0.000258622	–
	Var($\hat{\pi}$) Theoretical value	0.000311928	0.000299677	0.000276974	0.000256431	–
	Mean($\hat{\pi}$) Empirical value	0.2999511	0.2997295	0.2994269	0.2990468	–
0.5	Var($\hat{w}$) Empirical value	0.0010132	0.0010117	0.0010117	–	–
	Var($\hat{w}$) Theoretical value	0.001	0.001	0.001	–	–
	Mean($\hat{w}$) Empirical value	0.8999242	0.9001588	0.9001588	–	–
	Var($\hat{\pi}$) Empirical value	0.000317831	0.000302019	0.000280519	–	–
	Var($\hat{\pi}$) Theoretical value	0.000313408	0.000301267	0.000278768	–	–
	Mean($\hat{\pi}$) Empirical value	0.2999496	0.2995918	0.2990269	–	–
0.7	Var($\hat{w}$) Empirical value	0.0010132	0.0010117	–	–	–
	Var($\hat{w}$) Theoretical value	0.001	0.001	–	–	–
	Mean($\hat{w}$) Empirical value	0.8999242	0.9001588	–	–	–
	Var($\hat{\pi}$) Empirical value	0.000318655	0.000304912	–	–	–
	Var($\hat{\pi}$) Theoretical value	0.000315292	0.00030325	–	–	–
	Mean($\hat{\pi}$) Empirical value	0.2999515	0.2994566	–	–	–

3.2 *Simulation of $\hat{\mu}_X$ and $\hat{w}$ for Quantitative Model*

In this case also, the simulations results help validate our analytical findings. It may be noted from Table 2 that $\hat{\mu}_X$ and $\hat{w}$ are unbiased estimators for μ_X and w respectively. The theoretical and simulated variances of $\hat{\mu}_X$ gets closer as F increases for all values of T. For example, note from the Table 2 that for $T = 0.1$ and $F = 0.7$, the theoretical value of $Var(\hat{w}) = 0.001$ and simulated value of $Var(\hat{w}) = 0.0010035$ and theoretical value of $Var(\hat{\mu}_X) = 0.007998258$ and simulated value of $Var(\hat{\mu}_X) = 0.0073173$. The first order Taylor's Approximation was used to calculate the theoretical values for $Var(\hat{\mu}_X)$. Sihm et al. (2014) quantitative model is a one-stage model ($T = 0, F = 0$), thus, in Table 2 below that model is discussed only for one choice of (*T, F*) value, namely (0,0). In Table 2, (#) indicates that the corresponding quantity is for the Sihm et al. (2014) quantitative model.

Table 2: Simulation Results of Quantitative Model

Trials = 10000, $w = 0.9$, $\pi_a = 0.1$, $p_a = 0.5$, $p_b = 0.85$, $n = 1000$, $\mu_X = 2$ and $\mu_Y = 7$.						
F		T				
		0	0.1	0.3	0.5	0.7
0	Var($\hat{w}$) Empirical value	0.0010035	0.0010035	0.0010035	0.0010035	0.0010035
	Var($\hat{w}$) Empirical value (#)	0.001022				
	Var($\hat{w}$) Theoretical value	0.001	0.001	0.001	0.001	0.001
	Var($\hat{w}$) Theoretical value(#)	0.001				
	Mean($\hat{w}$) Empirical value	0.899673	0.899673	0.899673	0.899673	0.899673
	Mean($\hat{w}$) Empirical value(#)	0.900175				
	Var($\hat{\mu}_X$) Empirical value	0.0082023	0.0074183	0.0059583	0.0046236	0.0035064
	Var($\hat{\mu}_X$) Empirical value(#)	0.008129				
	Var($\hat{\mu}_X$) Theoretical value	0.016862408	0.014206666	0.009821285	0.006516794	0.0041341
	Var($\hat{\mu}_X$) Theoretical value(#)	0.008229				
	Mean($\hat{\mu}_X$) Empirical value	1.9994112	1.9999038	1.9999315	1.9999387	1.9998711
	Mean($\hat{\mu}_X$) Empirical value(#)	1.99962				
0.1	Var($\hat{w}$) Empirical value	0.0010035	0.0010035	0.0010035	0.0010035	0.0010035
	Var($\hat{w}$) Theoretical value	0.001	0.001	0.001	0.001	0.001
	Mean($\hat{w}$) Empirical value	0.899673	0.899673	0.899673	0.899673	0.899673
	Var($\hat{\mu}_X$) Empirical value	0.0081196	0.00737	0.0059227	0.0046329	0.0035291
	Var($\hat{\mu}_X$) Theoretical value	0.015178285	0.12899455	0.009102993	0.006198219	0.004057245
	Mean($\hat{\mu}_X$) Empirical value	1.999318	1.9998464	1.9998322	1.9999072	1.9999775
0.3	Var($\hat{w}$) Empirical value	0.0010035	0.0010035	0.0010035	0.0010035	–
	Var($\hat{w}$) Theoretical value	0.001	0.001	0.001	0.001	–
	Mean($\hat{w}$) Empirical value	0.899673	0.899673	0.899673	0.899673	–
	Var($\hat{\mu}_X$) Empirical value	0.008002	0.0072997	0.0059257	0.0046856	–
	Var($\hat{\mu}_X$) Theoretical value	0.012343024	0.010697987	0.007895423	0.005668409	–
	Mean($\hat{\mu}_X$) Empirical value	1.9993599	1.9997984	1.9995503	1.9998312	–

0.5	Var($\hat{w}$) Empirical value	0.0010035	0.0010035	0.0010035	–	–
	Var($\hat{w}$) Theoretical value	0.001	0.001	0.001	–	–
	Mean($\hat{w}$) Empirical value	0.899673	0.899673	0.899673	–	–
	Var($\hat{\mu}_X$) Empirical value	0.0079484	0.0072915	0.0059667	–	–
	Var($\hat{\mu}_X$) Theoretical value	0.010231069	0.009059284	0.007002568	–	–
	Mean($\hat{\mu}_X$) Empirical value	1.9990904	1.9998758	1.9996817	–	–
0.7	Var($\hat{w}$) Empirical value	0.0010035	0.0010035	–	–	–
	Var($\hat{w}$) Theoretical value	0.001	0.001	–	–	–
	Mean($\hat{w}$) Empirical value	0.899673	0.899673	–	–	–
	Var($\hat{\mu}_X$) Empirical value	0.0079748	0.0073173	–	–	–
	Var($\hat{\mu}_X$) Theoretical value	0.008857891	0.007998258	–	–	–
	Mean($\hat{\mu}_X$) Empirical value	1.9993201	1.9996364	–	–	–

4. Privacy Protection of Respondents

The aspect of privacy protection of respondents is an integral part of any RRT methodology. Lanke (1976), Giordano and Perri (2012), and Yan et al. (2009) have discussed this issue in detail. We examine this aspect for our proposed Three-Stage models. The Lanke (1976) privacy measure is used for the binary models while the Yan et al. (2009) privacy measure is used for the quantitative models. In Section 4.1, we briefly discuss the Lanke (1976) privacy measure and then tabulate the same for the Sihm et al. (2014) model and the proposed Three-Stage binary model. In Section 4.2, we recollect the Yan et al. (2009) privacy measure and then tabulate its values for the Sihm et al. (2014) model and the proposed Three-Stage quantitative model.

4.1 *Comparison of Binary Models using Lanke (1976) privacy measure*

Lanke (1976) defined a measure for privacy protection of respondents in binary RRT models. This privacy protection measure is based on the idea that higher the probability $P(X|R)$ of being classified in the group possessing *X* (the sensitive variable) by giving the response *R* (yes or no), the more stigmatizing it is to give that response. The Lanke measure is

$$L = \max\left[P(X|Yes), P(X|No)\right].$$

Thus, one method for randomized interviews may be considered to be more protective than another if $L = \max\left[P(X|Yes), P(X|No)\right]$ is smaller for the former method than for the latter.

Note that the Lanke (1976) privacy measure will take the value $P(X|Yes)$ when $P(X|Yes) > P(X|No)$. Let P_y denote the probability of 'yes' response. Let *X* denote the event that the respondent possesses the sensitive characteristic. Consider $P(X|Yes) - P(X|No)$,

$$P(X|Yes)-P(X|No)=\frac{P(Yes|X).P(X)}{P_y}-\frac{P(No|X).P(X)}{1-P_y}$$
$$=\frac{P(Yes|X)P(X)(1-P_y)-P(No|X)P(X)P_y}{P_y(1-P_y)}$$
$$=\frac{P(X)\{P(Yes|X)(1-P_y)-P(No|X)P_y\}}{P_y(1-P_y)}$$
$$=\frac{P(X)\{P(Yes|X)-[P(Yes|X)+P(No|X)]P_y\}}{P_y(1-P_y)}$$
$$=\frac{P(X)\{P(Yes|X)-P_y\}}{P_y(1-P_y)} \quad \text{since } [P(Yes|X)+P(No|X)]=1$$

But,

$$P(Yes|X)-P_y=(1-\pi_X)[P(Yes|X)-P(Yes|\bar{X})].$$

Thus,

$$P(X|Yes)-P(X|No)=\frac{P(X)\{P(Yes|X)-P_y\}}{P_y(1-P_y)}$$
$$=\frac{\pi_X(1-\pi_X)[P(Yes|X)-P(Yes|\bar{X})]}{P_y(1-P_y)}$$

Hence, the Lanke (1976) privacy measure will take the value $P(X|Yes)$ when $P(Yes|X)-P(Yes|\bar{X})>0$.

So for the binary models, we examine the difference between the above quantities.

Lanke (1976)'s Privacy Measure for Sihm et al. (2014) Binary Model

For Sihm et al. (2014) binary model,

$$P(Yes|X)-P(Yes|\bar{X})=(1-w)+w(p_b+(1-p_b)\pi_b)-w(1-p_b)\pi_b$$
$$=1-w+wp_b+w(1-p_b)\pi_b-w(1-p_b)\pi_b$$
$$=1-w+wp_b$$
$$=1-w(1-p_b)>0$$

As observed from the above discussion, $P(Yes|X)-P(Yes|\bar{X})>0$, so Lanke (1976) privacy measure will simply take the value $P(X|Yes)$ for the Sihm et al. (2014) model. We denote this value by L_1.

Thus,

$$L_1=P(X|Yes)=\frac{\pi[(1-w)+w\{p_b+(1-p_b)\pi_b\}]}{(1-w)\pi+w\{\pi\, p_b+(1-p_b)\pi_b\}}.$$

Lanke (1976)'s Privacy Measure for Three-Stage Binary Optional Model

For Three-Stage binary optional model,

$$\begin{aligned} P(Yes|X) - P(Yes|\overline{X}) &= T + F\{p_b + (1-p_b)\pi_b\} + (1-T-F)(1-w) \\ &\quad + (1-T-F)w\{p_b + (1-p_b)\pi_b\} - \{F(1-p_b)\pi_b - (1-T-F)w(1-p_b)\pi_b\} \\ &= T + Fp_b + (1-T-F)(1-w) + (1-T-F)wp_b > 0 \end{aligned}$$

As before, $P(Yes|X) - P(Yes|\overline{X}) > 0,$ so Lanke (1976) privacy measure will take the value $P(X|Yes)$ for Three-Stage binary optional model. We denote this value by L_2.

$$L_2 = P(X|Yes) = \frac{\pi[T + F[p_b + (1-p_b)\pi_b] + (1-T-F)[(1-w) + w\{p_b + (1-p_b)\pi_b\}]]}{T\pi + F\{p_b\pi + (1-p_b)\pi_b\} + (1-T-F)[(1-w)\pi + w\{p_b\pi + (1-p_b)\pi_b\}]}$$

Privacy Comparison between Sihm et al. (2014) Binary Model and Three-Stage Binary Optional Model

We compare the privacy measure of the Sihm et al. (2014) binary model with the Three-Stage optional binary model through Lanke (1976) measure. Comparing the two RRT models, the model with the smaller value of Lanke (1976) privacy measure will be more protective than the other. The comparison is summed up in the following result.

Theorem 3: *Three-Stage binary optional model offers more privacy than binary Sihm et al.* (2014) *model* $\Leftrightarrow$ $Tw < F(1-w)$.

Proof: Considering the difference of the privacy measures of the two models, we observe

$$\begin{aligned} L_2 - L_1 = &\frac{\pi[T + F[p_b + (1-p_b)\pi_b] + (1-T-F)[(1-w) + w\{p_b + (1-p_b)\pi_b\}]]}{T\pi + F\{p_b\pi + (1-p_b)\pi_b\} + (1-T-F)[(1-w)\pi + w\{p_b\pi + (1-p_b)\pi_b\}]} \\ &- \frac{\pi[(1-w) + w\{p_b + (1-p_b)\pi_b\}]}{(1-w)\pi + w\{\pi\, p_b + (1-p_b)\pi_b\}} \end{aligned}$$

Upon simplification, we get

$$L_2 - L_1 = \frac{\pi\pi_b(1-\pi)(1-p_b)(Tw - F(1-w))}{[(1-w)\pi + w\{\pi\, p_b + (1-p_b)\pi_b\}][T\pi + F\{p_b\pi + (1-p_b)\pi_b\} + (1-T-F)[(1-w)\pi + w\{p_b\pi + (1-p_b)\pi_b\}]]}$$

Since every term in the denominator is positive,

$$\begin{aligned} L_2 - L_1 < 0 &\Leftrightarrow \pi\pi_b(1-\pi)(1-p_b)(Tw - F(1-w)) < 0 \\ &\Leftrightarrow Tw < F(1-w) \end{aligned}$$

Q.E.D.

Observation: Consider a highly sensitive question for which $w = 0.9$, say, then it may be noted from Table 1 that for Sihm et al. (2014) binary model, the theoretical value for $Var(\hat{\pi}) = 0.00031$. But, if we instead take the value of $F = 0.7$ and $T = 0.05$, then $Tw < F(1-w)$ is also true for this combination of T, F and w. Thus, when the question is highly sensitive, and the Three−Stage binary model is used, then it certainly offers more privacy. More so, for Three−Stage binary model the theoretical value of $Var(\hat{\pi}) = 0.000309193$ for above combination of T and F. Hence, it may be concluded that once the parameters are chosen carefully, the Three-stage optional model offers better efficiency and more privacy than Sihm et al. (2014) model.

4.2 Comparison of Quantitative Models using Yan et al. (2009) privacy measure

For quantitative models also, the aspect of privacy protection of respondents is crucial for respondent co-operation. Yan et al. (2009) defined the measure of privacy protection as $\Delta = E[(Z - X)^2]$ where X is the true response of the sensitive variable and Z is the reported response. Thus, one model may be considered to be more protective than another if Δ for the former model is larger than the corresponding value for the latter.

Yan et al. (2009)'s Privacy Measure for Sihm et al. (2014) Quantitative Model

Let Δ_1 denote the Yan et al. (2009)'s privacy measure for Sihm et al. (2014) quantitative model.
Recall that for Sihm et al. (2014) quantitative model, the reported quantitative response is given by,

$$Z = \begin{cases} X \\ Y \end{cases} \text{with probability} \begin{cases} (1-w) + wp_b & \text{(sensitive question)} \\ w(1-p_b) & \text{(innocuous question)} \end{cases}$$

Then,

$$Z - X = \begin{cases} 0 \\ Y - X \end{cases} \text{with probability} \begin{cases} (1-w) + wp_b & \text{(sensitive question)} \\ w(1-p_b) & \text{(innocuous question)} \end{cases}$$

Assuming X and Y to be independent random variables, we get

$$\begin{aligned} \Delta_1 &= E[(Z-X)^2] \\ &= E[(Y-X)^2]w(1-p_b) \\ &= E(Y^2 + X^2 - 2XY)w(1-p_b) \\ &= [E(Y^2) + E(X^2) - 2E(X)E(Y)]w(1-p_b) \\ &= [\sigma_X^2 + \mu_X^2 + \sigma_Y^2 + \mu_Y^2 - 2\mu_X\mu_Y]w(1-p_b) \\ &= [\sigma_X^2 + \sigma_Y^2 + (\mu_X - \mu_Y)^2]w(1-p_b) \end{aligned}$$

Yan et al. (2009)'s Privacy Measure for Three-Stage Quantitative Optional Model

Let Δ_2 denote the Yan et al. (2009)'s privacy measure for the proposed Three-Stage quantitative optional model.

For Three-Stage quantitative optional model, the reported quantitative response is

$$Z = \begin{cases} X \\ Y \end{cases} \text{with probability} \begin{cases} T + Fp_b + (1-T-F)[(1-w) + wp_b] & \text{(sensitive question)} \\ F(1-p_b) + (1-T-F)w(1-p_b) & \text{(innocuous question)} \end{cases}$$

Then,

$$Z-X=\begin{cases}0\\ Y-X\end{cases}\text{with probability}\begin{cases}T+Fp_b+(1-T-F)\left[(1-w)+wp_b\right] & \text{(sensitive question)}\\ F(1-p_b)+(1-T-F)w(1-p_b) & \text{(innocuous question)}\end{cases}$$

Assuming X and Y to be independent random variables, we get,

$$\begin{aligned}\Delta_2&=E[(Z-X)^2]\\&=E[(Y-X)^2]\{F(1-p_b)+(1-T-F)w(1-p_b)\}\\&=E(Y^2+X^2\!-\!2XY)\{F(1-p_b)+(1-T-F)w(1-p_b)\}\\&=\left(E(Y^2)+E(X^2)-2E(X)E(Y)\right)\{F(1-p_b)+(1-T-F)w(1-p_b)\}\\&=[\sigma_X^2+\mu_X^2+\sigma_Y^2+\mu_Y^2-2\mu_X\mu_Y]\{F(1-p_b)+(1-T-F)w(1-p_b)\}\\&=[\sigma_X^2+\sigma_Y^2+(\mu_X-\mu_Y)^2](1-p_b)\{F+(1-T-F)w\}\end{aligned}$$

Privacy Comparison between Sihm et al. (2014) model and Three-Stage Model (both quantitative)

As done earlier for binary models, we compare the privacy measure of the two quantitative models using Yan et al. (2009) privacy measure. The findings are given below:

Theorem 4: *Three-Stage quantitative model offers more privacy than Sihm et al.* (2014) *quantitative model* $\Leftrightarrow Tw<F(1-w)$.

Proof: Considering the difference of the privacy measures of the models, we observe

$$\begin{aligned}\Delta_2-\Delta_1&=[\sigma_X^2+\sigma_Y^2+(\mu_X-\mu_Y)^2](1-p_b)\{F+(1-T-F)w\}-[\sigma_X^2+\sigma_Y^2+(\mu_X-\mu_Y)^2](1-p_b)w\\&=[\sigma_X^2+\sigma_Y^2+(\mu_X-\mu_Y)^2](1-p_b)\{F+(1-T-F)w-w\}\\&=[\sigma_X^2+\sigma_Y^2+(\mu_X-\mu_Y)^2](1-p_b)\{F(1-w)-Tw\}\end{aligned}$$

$$\Delta_2-\Delta_1>0\Leftrightarrow F(1-w)>Tw$$

Q.E.D.

Observation:

- It may be observed that the condition obtained in Theorem 3 and 4 are same, that is, when the parameters T, F and w are chosen such that $Tw<F(1-w)$ is satisfied, then the Three−Stage model (both binary and quantitative) offers more privacy to the respondents as compared to the corresponding Sihm et al. (2014) model.
- As observed in case of binary models, we again consider a highly sensitive question for which $w=0.9$, say, from Table 2 we get that for Sihm et al. (2014) quantitative model, the theoretical value for $Var(\hat{\mu}_X)=0.008229$. If Three−Stage quantitative model is used instead with $F=0.7$ and $T=0.05$, then $Tw<F(1-w)$ is true for this combination of T, F and w and so, from Theorem 4, the Three−Stage

quantitative model offers more privacy to the respondents as compared to the quantitative Sihm et al. (2014) model. More so, for Three − Stage quantitative model the theoretical value of $Var(\hat{\mu}_X) = 0.0082211$ for above combination of *T* and *F*. Hence, it may be concluded that if the parameters are chosen carefully, the Three−Stage quantitative optional model offers better efficiency and more privacy than Sihm et al. (2014) quantitative model.

5. Conclusion

In this paper, we propose three-stage versions of the Sihm et al. (2014) modified optional unrelated question RRT models for both binary and quantitative response situations. When the parameters *T* and *F* of the Three-stage model are chosen to be 0, then this model reduces to the corresponding Sihm et al. (2014) model. The simulation study shows that for appropriate choices of (*T, F*), the proposed models work better than the corresponding Sihm et al. (2014) model. We also observe that the Three−Stage optional models offer more privacy than corresponding Sihm et al. (2014) model $\Leftrightarrow Tw < F(1-w)$. So, depending upon the sensitivity level of the underlying research question, one can find values of T and F for the Three−Stage model so that respondents can be offered more privacy along with the a lower value of the variance of $\hat{\pi}$ and $\hat{\mu}_X$ (depending upon the response scenario) with respect to the corresponding Sihm et al. (2014) model.

References

1. Giordano, S., Perri, P.F., (2012). Efficiency Comparison of Unrelated Question Models Based on Same Privacy Protection Degree, *Statistical Papers*, 53, 987-999.
2. Greenberg, B.G., Abul-Ela, A. L. A., Simmons, W. R., and Horvitz, D. G. (1969). The Unrelated Question Randomized Response Model: Theoretical Framework, *Journal of the American Statistical Association*, 64, 520-529.
3. Greenberg, B. G., Kuebler, R. R., Abernathy, J. R., & Horvitz, D. G. (1971). Application of the randomized response technique in obtaining quantitative data, *Journal of the American Statistical Association*, 66 (334), 243-250.
4. Gupta, S., Gupta, B., and Singh, S. (2002). Estimation of sensitivity level of personal interview survey question, *Journal of Statistical Planning and Inference*, 100, 239-247.
5. Gupta, S., Tuck, A., Spears Gill, T., and Crowe, M., (2013). Optional Unrelated Question Randomized Response Models, *Involve: A Journal of Mathematics*, 6(4), 483-492.
6. Lanke, J., (1976). On the Degree of Protection in Randomized Interviews, *International Statistical Reviews*, 44 (2), 197-203.
7. Mangat, N. S., and Singh R., (1990). An Alternative Randomized Response Procedure, *Biometrika*, 77(2), 439-442.
8. Mehta, S., Dass, B. K., Shabbir, J., and Gupta S. N., (2012). A Three-Stage Optional Randomized Response Model, *Journal of Statistical Theory and Practice*, 6(3), 417-427.
9. Sihm, J. S., Chhabra A., and Gupta S. N., (2014). An Optional Unrelated Question RRT Model, (Forthcoming in *Involve: A Journal of Mathematics*).
10. Warner, S. L., (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias, *Journal of American Statistical Association*, 60 (309), 63-69.
11. Yan, Z., Wang, J. and Lai, J. (2009). An efficiency and protection degree-based comparison among the quantitative randomized response strategies, *Communications in Statistics − Theory and Methods*, 38, 400-408.