Published in final edited form as:

Stat Methods Appt. 2016 November ; 25(4): 601–621. doi:10.1007/s10260-016-0351-1.

An alternative to unrelated randomized response techniques with logistic regression analysis

Shu-Hui Hsieh¹, Shen-Ming Lee², Chin-Shang Li³, and Su-Hao Tu¹

¹Center for Survey Research, Research Center for Humanities and Social Sciences, Academia Sinica, Taipei, Taiwan

²Department of Statistics, Feng Chia University, Taichung, Taiwan

³Division of Biostatistics, Department of Public Health Sciences, University of California, Davis, CA, USA

Abstract

The randomized response technique (RRT) is an important tool that is commonly used to protect a respondent's privacy and avoid biased answers in surveys on sensitive issues. In this work, we consider the joint use of the unrelated-question RRT of Greenberg et al. (J Am Stat Assoc 64:520–539, 1969) and the related-question RRT of Warner (J Am Stat Assoc 60:63–69, 1965) dealing with the issue of an innocuous question from the unrelated-question RRT. Unlike the existing unrelated-question RRT of Greenberg et al. (1969), the approach can provide more information on the innocuous question by using the related-question RRT of Warner (1965) to effectively improve the efficiency of the maximum likelihood estimator of Scheers and Dayton (J Am Stat Assoc 83:969–974, 1988). We can then estimate the prevalence of the sensitive characteristic by using logistic regression. In this new design, we propose the transformation method and provide large-sample properties. From the case of two survey studies, an extramarital relationship study and a cable TV study, we develop the joint conditional likelihood method. As part of this research, we conduct a simulation study of the relative efficiencies of the proposed methods. Furthermore, we use the two survey studies to compare the analysis results under different scenarios.

Keywords

Randomized response technique; Transformation method; Joint conditional likelihood method

1 Introduction

Having accurate answers to surveys on social issues is important, especially when questions are related to a respondent's privacy. Respondents frequently have incentives to not tell the truth when questions touch upon moral, legal, or other sensitive issues. If we ask sensitive questions, and ignore the possibility that the respondents will not tell the truth, this will cause estimation error and bias. Therefore, methods have been developed to minimize the likelihood of error and bias. The randomized response technique (RRT) proposed by Warner

(1965) was the first attempt to obtain reliable information for estimating the proportion of a sensitive attribute in a population without revealing the respondent's actual status. Different modifications of Warner's RRT (1965) have been developed by various authors, such as Greenberg et al. (1969), Horvitz et al. (1967), Moors (1971), Raghavarao (1978), Mangat and Singh (1990), Kuk (1990), Mangat (1994), Singh et al. (2000), Haung (2004), Kim and Warde (2004), Chang et al. (2004), and Gjestvang and Singh (2006). In the RRT, a respondent is willing to answer and tell the truth to sensitive questions through some random devices (for instance, dice, playing cards, or coins), while the interviewer is blind to the outcome.

In sex research, where the percentages of extramarital sex are often underestimated, a bias may occur due to the method of data collection, assessment methods (such as questionnaire designs), and estimation models. Among these issues related to reporting of sexual behavior, a crucial concern is the tendency to under-report based on information gathered from retrospective self-report interviews (Gribble et al. 1999). Moreover, a meta-analysis by Lensvelt-Mulders et al. (2005) showed that the RRT yields more valid prevalence estimates than other methods for sensitive questions. For example, the Taiwan Social Change Survey (TSCS), conducted in 2012 on Taiwanese aged 18 years old or older by computer-assisted personal interviewing, investigated the rate of people in extramarital relationships in Taiwan by using the unrelated-question RRT of Greenberg et al. (1969). The question set was designed as follows:

A: Have you ever had sex with someone other than your spouse/partner?

B: Were you born in the month of January, February, or March?

Under this RRT, the respondent determined which question to answer based on what card she/he picked from a well-shuffled deck of 40 numbered playing cards, while the interviewer was blind to the outcome. Note that the values of playing cards are from 1 to 5 with probabilities 0.2, 0.1, 0.2, 0.4, and 0.1, respectively. The respondent was then requested to answer question A if the card value was 1, 2, or 3; otherwise she/he answered question B. We incorporated an unrelated question or an innocuous question (such as question B) to make the respondent believe that it was safe to answer the question because her/his card value was not revealed to the interviewer. Moreover, when the respondent is willing to answer the unrelated question of RRT, she/he might be more willing to cooperate with the interviewer because she/he believes that her/his privacy is protected. Therefore, the estimate will be more accurate.

The most important RRT claim is that it yields more valid point estimates of sensitive behavior. It is important to note that even though sensitive behavior is measured by using the RRT, it is still possible to link sensitive behavior to demographic variables with a specially adapted version of the logistic regression. Based on the procedures of Warner (1965), Greenberg et al. (1969), Scheers and Dayton (1988) were the first to present a covariate-randomized-response model to obtain estimators. Corstange (2004) proposed a method to estimate the parameters of a hidden logistic regression model. Hsieh et al. (2009) proposed two semiparametric approaches to estimate the parameters of a logistic regression model with missing covariates for an RRT model. Kim and Heo (2013) combined the two most

celebrated randomized response techniques, which are the related-question RRT of Warner (1965) and the unrelated-question RRT of Greenberg et al. (1969) with group testing. Magder and Hughes (1997) discussed a logistic regression model where the response variable is subject to misclassification comparable to the perturbation induced by the RRT. Van den Hout et al. (2007) discussed univariate and multivariate logistic regression models when the response variables are subject to randomized response.

Our work is motivated by the case of the two existing survey study data examples, an extramarital relationship study from the TSCS and a cable TV study on unauthorized cable TV connection in Taiwan by Hsieh et al. (2009), in which answers of an innocuous question from the unrelated-question RRT can be indirectly and completely observed from answers of a demographic question that is related to the innocuous question. To improve the protection of privacy, we consider the joint use of the unrelated-question RRT of Greenberg et al. (1969) and the related-question RRT of Warner (1965) dealing with the issue of an innocuous question from the unrelated-question RRT. Unlike the existing unrelated-question RRT of Greenberg et al. (1969), the approach can provide more information on the innocuous question by using the related-question RRT of Warner (1965) to effectively improve the efficiency of the maximum likelihood estimator of Scheers and Dayton (1988). We can then estimate the prevalence of the sensitive characteristic by using logistic regression. The estimation of this prevalence is key to estimating the population proportion that will answer question A a certain way. To solve this problem, we propose a new method, called a transformation method, which uses available information from survey answers. Furthermore, if answers to the innocuous question can be completely observed, we can specifically develop the joint conditional likelihood method based on the missing data framework from the cross table to analyze data. Finally, we compare the results of all the methods to explore the advantages and disadvantages of each method.

The remainder of this paper is organized as follows. In Sect. 2, we briefly review the unrelated-question RRT method of Greenberg et al. (1969) within a logistic regression model framework of Scheers and Dayton (1988). In Sect. 3, based on the logistic regression model, we present the transformation method and the asymptotic properties. In Sect. 4, we discuss the case of the extramarital relationship study and the cable TV study. In these two existing studies answers to the innocuous question can be indirectly and completely observed, and we develop the joint conditional likelihood method based on the missing data framework. This method uses a cross table in which a few cases are indirectly elicited to provide answers to the sensitive question. Sect. 5 conducts a simulation study to assess the performance of our estimators and to compare them with the maximum likelihood estimator of Scheers and Dayton (1988). In Sect. 6, in order to support the simulation results, we use the two real data examples to compare the results under different scenarios. Finally, we conclude with a discussion and remarks in Sect. 7.

2 Review of the unrelated-question RRT method

We briefly review the unrelated-question RRT method of Greenberg et al. (1969) within the logistic regression model framework of Scheers and Dayton (1988). Under the survey from TSCS of 2012, we let Y and T denote the of binary outcome variables corresponding,

respectively, to answers of sensitive question A and innocuous question B. Under the unrelated-question RRT of Greenberg et al. (1969), Y cannot be observed. Instead, we can only observe a binary response variable Y^0 , which corresponds to either question A or B. Question A or B for each person is provided by a latent binary variable based on chance game Q with probability $P(Q = 1) = p$ and $P(Q = 0) = 1 - p$. Therefore, we have $Y^0 = Y$ with probability p and $Y^0 = T$ with probability $1 - p$. Moreover, Y is a latent binary RRT variable corresponding to the answer to sensitive question A in which $Y = 1$ if the answer is “Yes”, and $Y = 0$ if the answer is “No”. Let X be a covariate vector, which is always observed. We consider the following logistic regression model:

$$P(Y = 1 | X) = H(\beta_0 + \beta_1^t X) = H(\beta^t \mathcal{X}), \quad (1)$$

where t is the transpose operator, $H(u) = \{1 + \exp(-u)\}^{-1}$, $\mathcal{X} = (1, X^t)^t$, and $\beta = (\beta_0, \beta_1^t)^t$ is a vector unknown parameters.

Let n be the sample size. Under the unrelated-question RRT of Greenberg et al. (1969), we obtain (Y_i^0, X_i) , $i = 1, 2, \dots, n$. Based on model (1) and letting $P(T_i = 1 | X_i) = c$, we can have

$$\begin{aligned} P(Y_i^0 = 1 | X_i) &= P(Y_i^0 = 1 | Q_i = 1, X_i)P(Q_i = 1) + P(Y_i^0 = 1 | Q_i = 0, X_i)P(Q_i = 0) \\ &= pP(Y_i = 1 | X_i) + (1 - p)P(T_i = 1 | X_i) \\ &= pH(\beta^t \mathcal{X}_i) + c(1 - p). \end{aligned} \quad (2)$$

c in model (2) denotes the proportion of the population having the question B in which $T = 1$ if the answer is “Yes”. Greenberg et al. (1969) dealt with the two situations where c is known and it is unknown. If c is unknown, then two sub-samples of size n_1 and n_2 such that $n = n_1 + n_2$ are required; otherwise a single sample of size n is required.

Under model (2) and c is known, the likelihood is given (see Scheers and Dayton 1988) as follows:

$$\begin{aligned} L(\beta) &= \prod_{i=1}^n P(Y_i^0 = 1 | X_i)^{Y_i^0} P(Y_i^0 = 0 | X_i)^{1 - Y_i^0} \\ &= \prod_{i=1}^n [pH(\beta^t \mathcal{X}_i) + c(1 - p)]^{Y_i^0} [1 - pH(\beta^t \mathcal{X}_i) - c(1 - p)]^{1 - Y_i^0}. \end{aligned}$$

Hence we define the following score function:

$$\begin{aligned}
 U_{M,n}(\boldsymbol{\beta}) &= \frac{1}{\sqrt{n}} \frac{\partial}{\partial \boldsymbol{\beta}} \ln L(\boldsymbol{\beta}) \\
 &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathcal{X}_i A_i(\boldsymbol{\beta}) \{Y_i^0 - [pH(\boldsymbol{\beta}^t \mathcal{X}_i) + (1-p)c]\}, \quad (3)
 \end{aligned}$$

where

$$A_i(\boldsymbol{\beta}) = \frac{pH^{(1)}(\boldsymbol{\beta}^t \mathcal{X}_i)}{[pH(\boldsymbol{\beta}^t \mathcal{X}_i) + c(1-p)][1 - [pH(\boldsymbol{\beta}^t \mathcal{X}_i) + c(1-p)]]}$$

for $H^{(1)}(\boldsymbol{\beta}^t \mathcal{X}_i) = H(\boldsymbol{\beta}^t \mathcal{X}_i)[1 - H(\boldsymbol{\beta}^t \mathcal{X}_i)]$. The maximum likelihood estimator of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}_M$, can be obtained by solving $U_{M,n}(\boldsymbol{\beta}) = \mathbf{0}$ with the Newton–Raphson method. Note that when $p = 1$, all the data are from the direct question ($Y_i^0 = Y_i$) model; hence this reduces to a logistic regression model. But the direct question about sensitive items often yields a non-response or false response. Recent meta-analyses have shown that the RRT methods can outperform significantly more direct ways of asking sensitive questions (Lensvelt-Mulders et al. 2005).

3 Proposed estimation method

In this work, we consider, as an alternative to the unrelated RRT, the joint use of the unrelated-question RRT of Greenberg et al. (1969) and the related-question RRT of Warner (1965) dealing with the issue of an innocuous question from the unrelated-question RRT. The approach can provide additional information on the innocuous question for developing a more efficient estimation method. For example, the related-question set for innocuous question B from the unrelated-question RRT was designed as follows:

B: Were you born in the month of January, February, or March?

B^c: Were you not born in the month of January, February, or March?

Here question B is a basic demographic question, and B^c is a complementary question to question B. A randomization device (for instance, chance game, a draw from playing cards or the roll of a dice) is used to decide which of the two questions is answered with “Yes” or “No”. Under this RRT, the answer to the innocuous question B depends partly on the respondent’s true status and partly on the outcome of a randomizing device. Hence we can only observe a binary response variable T^0 , which corresponds to either question B or B^c. Question B or B^c for each person is provided by a latent binary variable based on the chance game with probability π and $1 - \pi$. Thus, we have $T^0 = T$ with probability π and $T^0 = 1 - T$ with probability $1 - \pi$.

Under the joint use of the unrelated-question–RRT of Greenberg et al. (1969) and the related-question RRT of Warner (1965) dealing with the issue of an innocuous question from

the unrelated-question RRT, we first used playing cards to ask basic demographic questions by the related-question RRT. Next, we used another playing card method to get the information on extramarital relationships by the unrelated-question RRT. We would tell interviewees that both techniques can provide useful information and protect their privacy. In Fig. 1, we show the probability mass diagrams for the values of T_i^0 by the selection of T_i or $1 - T_i$ from the related-question RRT and the values of Y_i^0 by the selection of Y_i or T_i from the unrelated-question RRT, respectively. The observed data set is (Y_i^0, X_i, T_i^0) , $i = 1, 2, \dots, n$. It is, however, noted that under the classic RRT only (Y_i^0, X_i) , $i = 1, \dots, n$, can be observed. Moreover, based on model (1), we propose a transformation method that uses available information from the survey question set.

3.1 Transformation method

In this section we use a transformation score technique to develop the transformation method for estimating the parameters of the logistic regression model. Let (Y_i^0, X_i, T_i^0) , $i = 1, 2, \dots, n$, be the data collected by the joint use of the unrelated-question RRT of Greenberg et al. (1969) and the related-question RRT of Warner (1965) dealing with the issue of an innocuous question from the unrelated-question RRT. The joint probability mass function of (Y_i^0, T_i^0) given X_i can be derived as

$$P(Y_i^0 = 0, T_i^0 = 0 | X_i) = \pi \{ p(1 - H(\beta^t X_i))(1 - c) + (1 - p)(1 - c) \} + (1 - \pi) p(1 - H(\beta^t X_i))_c$$

$$P(Y_i^0 = 0, T_i^0 = 1 | X_i) = \pi \{ p(1 - H(\beta^t X_i))c \} + (1 - \pi) \{ p(1 - H(\beta^t X_i))(1 - c) + (1 - p)(1 - c) \}$$

$$P(Y_i^0 = 1, T_i^0 = 0 | X_i) = \pi \{ pH(\beta^t X_i)(1 - c) \} + (1 - \pi) \{ pH(\beta^t X_i)c + (1 - p)c \} \text{ and}$$

$P(Y_i^0 = 1, T_i^0 = 1 | X_i) = \pi \{ pH(\beta^t X_i)c + (1 - p)c \} + (1 - \pi) \{ pH(\beta^t X_i)(1 - c) \}$. Because, as can be seen from the above expressions, the estimation method based on the joint likelihood is complicated, the data needs to be translated into the best estimate of the quantity of interest (González and Davier 2013). Note that the Y_i^0 can directly provide information on estimation of β , however, the T_i^0 cannot directly provide information on that because β is not involved in the probability $P(T_i^0 = 1) = \pi c + (1 - \pi)(1 - c)$. The key idea of proposed estimation method is partition the information of Y_i^0 by $T_i^0 = 0$ and $T_i^0 = 1$ so that the Y_i^0 can provide disjoint information on estimation of β . Therefore, we translate the two variables Y_i^0 and T_i^0 to $Z_i = Y_i^0 T_i^0$ and $S_i = Y_i^0(1 - T_i^0)$ to attain this purpose. It can then be shown that

$$\begin{aligned}
P(Z_i = 1 | X_i) &= P(Y_i^0 T_i^0 = 1 | X_i) \\
&= \pi P(Y_i^0 = 1, T_i = 1 | X_i) + (1 - \pi) P(Y_i^0 = 1, T_i = 0 | X_i) \\
&= \pi [pH(\beta^t \mathcal{X}_i)^c + (1 - p)c] + (1 - \pi) [pH(\beta^t \mathcal{X}_i)(1 - c)] \\
&\equiv H_Z(X_i; \beta),
\end{aligned}$$

and

$$\begin{aligned}
P(S_i = 1 | X_i) &= P(Y_i^0(1 - T_i^0) = 1 | X_i) \\
&= P(Y_i^0 = 1, T_i^0 = 0 | X_i) \\
&= \pi P(Y_i^0 = 1, T_i = 0 | X_i) + (1 - \pi) P(Y_i^0 = 1, T_i = 1 | X_i) \\
&= \pi [pH(\beta^t \mathcal{X}_i)(1 - c)] + (1 - \pi) [pH(\beta^t \mathcal{X}_i)^c + (1 - p)c] \\
&\equiv H_S(X_i; \beta).
\end{aligned}$$

To estimate β , we define the variance of Z_i and S_i as follows:

$$V(X_i, \beta) = \text{var} \begin{pmatrix} Z_i \\ S_i \end{pmatrix} = \begin{bmatrix} H_Z^{(1)}(X_i, \beta) & -H_Z(X_i, \beta)H_S(X_i, \beta) \\ -H_Z(X_i, \beta)H_S(X_i, \beta) & H_S^{(1)}(X_i, \beta) \end{bmatrix},$$

where $H_Z^{(1)}(X_i, \beta) = H_Z(X_i, \beta)[1 - H_Z(X_i, \beta)]$ and $H_S^{(1)}(X_i, \beta) = H_S(X_i, \beta)[1 - H_S(X_i, \beta)]$.

Based on Z_i , S_i and $V(X_i, \beta)$, the estimating score function is defined as follows:

$$\begin{aligned}
U_{T,n}(\beta) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\frac{\partial H_Z(X_i, \beta)}{\partial \beta}, \frac{\partial H_S(X_i, \beta)}{\partial \beta} \right) (V(X_i, \beta))^{-1} \begin{pmatrix} Z_i - H_Z(X_i, \beta) \\ S_i - H_S(X_i, \beta) \end{pmatrix} \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \{ \mathcal{X}_i K_i(\beta) [Z_i - H_Z(X_i, \beta)] + \mathcal{X}_i \lambda_i(\beta) [S_i - H_S(X_i, \beta)] \},
\end{aligned} \tag{4}$$

where

$$\begin{aligned}
K_i(\beta) &= H^{(1)}(\beta^t \mathcal{X}_i) H_S(X_i, \beta) \\
&\times \left\{ \frac{[\pi c + (1 - \pi)(1 - c)](1 - H_S(X_i, \beta)) + [\pi(1 - c) + (1 - \pi)c]H_Z(X_i, \beta)}{H_Z^{(1)}(X_i, \beta)H_S^{(1)}(X_i, \beta) - [H_Z(X_i, \beta)H_S(X_i, \beta)]^2} \right\},
\end{aligned}$$

and

$$\lambda_i(\beta) = H^{(1)}(\beta^t \mathcal{X}_i) H_Z(X_i; \beta) \\ \times \left\{ \frac{[\pi c + (1 - \pi)(1 - c)] H_S(X_i; \beta) + [\pi(1 - c) + (1 - \pi)c] (1 - H_Z(X_i; \beta))}{H_Z^{(1)}(X_i; \beta) H_S^{(1)}(X_i; \beta) - [H_Z(X_i; \beta) H_S(X_i; \beta)]^2} \right\}.$$

The transformation method estimator of β , denoted by $\hat{\beta}_T$, is the root of $U_{T,n}(\beta) = \mathbf{0}$, which can be obtained by the Newton-Raphson method.

3.2 Asymptotic result

We now provide the asymptotic results of the proposed estimator $\hat{\beta}_T$. Here the estimating score function $U_{T,n}(\beta)$ in (4) is the sum of $S_i(\beta)$ s divided by $\sqrt{n}$, where $S_i(\beta) = \{\mathcal{X}_i K_i(\beta) [Z_i - H_Z(X_i; \beta)] + \mathcal{X}_i \lambda_i(\beta) [S_i - H_S(X_i; \beta)]\}$. We use a Taylor's series expansion of the score function to show that the estimator $\hat{\beta}_T$ of β is consistent. We consider $G_{T,n}(\beta) = \frac{-1}{\sqrt{n}} \frac{\partial U_{T,n}(\beta)}{\partial \beta}$ to show that $G_{T,n}(\beta) \xrightarrow{P} G_T(\beta) = E[G_{T,n}(\beta)]$ is uniform in a neighborhood of the true value of β by the strong law of large numbers. It can be proved that a unique consistent solution exists for the estimating equations $U_{T,n}(\beta) = \mathbf{0}$ in a neighborhood of the true value of β by the Inverse Function Theorem of Foutz (1977). Therefore, $\hat{\beta}_T$ is shown to be a consistent estimator of β .

To derive the asymptotic distribution of $\sqrt{n}(\hat{\beta}_T - \beta)$, we use a Taylor's series expansion of $U_{T,n}(\hat{\beta}_T)$ at β to have

$$\mathbf{0} = U_{T,n}(\hat{\beta}_T) = U_{T,n}(\beta) + \frac{\partial U_{T,n}(\beta)}{\partial \beta} (\hat{\beta}_T - \beta) + o_p(1) \\ = U_{T,n}(\beta) - G_{T,n}(\beta) \sqrt{n} (\hat{\beta}_T - \beta) + o_p(1).$$

Because $G_{T,n}(\beta) \xrightarrow{P} G_T(\beta)$ and $U_{T,n}(\beta) \xrightarrow{d} N(\mathbf{0}, Q_T(\beta))$, where $Q_T(\beta) = \text{Cov}[U_{T,n}(\beta)]$, it can be shown that $\sqrt{n}(\hat{\beta}_T - \beta)$ is asymptotically normally distributed with mean $\mathbf{0}$ and covariance matrix

$$\Delta_T = G_T^{-1}(\beta) Q_T(\beta) [G_T^{-1}(\beta)]^t.$$

In practice, a consistent estimator of Δ_T is needed. For any vector a , define $a^{\otimes 2} = aa^t$. Assume that $G_{T,n}(\hat{\beta}_T)$ and $Q_{T,n}(\hat{\beta}_T)$ are the estimators of $G_T(\beta)$ and $Q_T(\beta)$, respectively, where

$$G_{T,n}(\hat{\beta}_T) = \frac{-1}{\sqrt{n}} \frac{\partial U_{T,n}(\beta)}{\partial \beta} \Big|_{\beta = \hat{\beta}_T}$$

and

$$Q_{T,n}(\hat{\beta}_T) = \frac{1}{n} \sum_{i=1}^n [s_{T,i}(\hat{\beta}_T)]^{\otimes 2}.$$

Here each component of $G_{T,n}(\cdot)$ is a negative mean of independent and identically distributed random variables. We can then show that $G_{T,n}(\hat{\beta}_T) - G_{T,n}(\beta) \xrightarrow{P} 0$ because $\hat{\beta}_T \xrightarrow{P} \beta$. Using the weak law of large numbers, it can be shown that $G_{T,n}(\beta) \xrightarrow{P} G_T(\beta)$. Consequently, using Slutsky's theorem can justify $G_{T,n}(\hat{\beta}_T) \xrightarrow{P} G_T(\beta)$. The same arguments can be used to show $Q_{T,n}(\hat{\beta}_T) \xrightarrow{P} Q_T(\beta)$ and have a consistent estimator of Δ_T given by

$$\hat{\Delta}_T = G_{T,n}^{-1}(\hat{\beta}_T) Q_{T,n}(\hat{\beta}_T) [G_{T,n}^{-1}(\hat{\beta}_T)]^t.$$

As before, we can also provide the asymptotic results of $\hat{\beta}_M$. Here the estimating equation in (3) is the sum of $S_i(\beta)$ s divided by $\sqrt{n}$, so let $S_i(\beta) = \mathcal{X}_i^t A_i(\beta) \{Y_i^0 - [pH(\beta^t \mathcal{X}_i) + (1-p)c]\}$. It can then be shown that $\sqrt{n}(\hat{\beta}_M - \beta)$ is asymptotically normally distributed with mean $\mathbf{0}$ and covariance matrix Δ_M .

4 A special case of the proposed method and design consideration

In this section, we discuss the special case where, when $\pi = 1$, answers of the innocuous question B from the unrelated-question RRT can be obtained directly. Moreover, for the two existing study examples, the extramarital relationship study and cable TV study, an answer to the innocuous question B from the unrelated-question RRT can be obtained indirectly from a demographic question that is related to the innocuous question B. First, practical issues are raised by an existing example from the TSCS in 2012. Depending on the sample and the reason for the survey, this could be viewed as an identifying demographic question as follows:

C: When were you born?

The question C is related to the innocuous question B at the beginning of the TSCS. It is a very basic demographic question that is not sensitive, and the respondent will probably be willing to provide an answer.

Next, data were collected to study the proportion of unauthorized cable TV users in Taichung, Yunlin, and Taipei, Taiwan in 2004, and were analyzed by Hsieh et al. (2009). The respondents were required to answer one of the questions of unrelated-question RRT below based on the last digit of their identification number.

A': Is your cable TV connection unauthorized?

B': Is the number of family members living with you an odd number?

Here we could view a demographic question in the same survey as follows:

C': Including you, how many people live in your house?

The answer to question C' could provide direct information on the innocuous question B'. This is not deliberately designed *a priori* to collect the answer of the innocuous question in order to improve the efficiency of estimation.

Under the assumption of truthful response by all respondents, question C or C' can provide valuable information because it can also cause answers of the innocuous question B or B' to be indirectly and completely observed after the randomized response data are collected. As before, we can also use the estimating score function in (4) with $\pi = 1$ to show that the transformation method estimator is a consistent estimator based on (Y_i^0, X_i, T_i) . Moreover, we specifically develop the joint conditional likelihood method based on the missing data framework from the cross table of (Y_i^0, T_i) . Note that this is not deliberately designed *a priori* to collect the value of T_i in order to improve the efficiency of estimation. In addition, this information of T_i only improves the efficiency of the estimation for inference, but does not disturb privacy of individuals.

4.1 Joint conditional likelihood method

The joint conditional likelihood method is valid only when T_i is completely observed or $\pi = 1$. Based on the fact that T_i is completely observed, we can elicit a few values of the answer to the sensitive question Y_i between Y_i^0 and T_i based on the respondent's truthful report. The joint conditional likelihood method needs some valid Y_i s, and this depends on the pairs $\{Y_i^0 = 1, T_i = 0\}$ and $\{Y_i^0 = 0, T_i = 1\}$. Note that the value of Y_i^0 is equal to Y_i or T_i . If Y_i^0 is not equal to T_i , this implies that the value of Y_i^0 is from Y_i , i.e., $Y_i^0 = Y_i$. Hence the $\{Y_i^0 = 1, T_i = 0\}$ and $\{Y_i^0 = 0, T_i = 1\}$ can infer that $Y_i = 1$ and $Y_i = 0$, respectively. Therefore, it is possible to obtain a few answers to the sensitive question Y_i . The value of Y_i , however, cannot be ascertained when $Y_i^0 = T_i$. Moreover, based on data on a few sensitive characteristics, the complete-case estimator has two potential disadvantages: (a) loss of efficiency and (b) the potential to yield inconsistent estimates, when a data set is not a random subsample of the original case data set.

For the sake of illustration let δ_i indicate whether Y_i can be elicited ($\delta_i = 1$) or not elicited ($\delta_i = 0$). Therefore, the validation data set ($\delta_i = 1$) consists of (Y_i^0, X_i, T_i, Y_i) , and the non-validation data set ($\delta_i = 0$) consists of (Y_i^0, X_i, T_i) . Based on model (1), the selection probabilities are given by

$$\begin{aligned} P(\delta_i = 1 | X_i) &= P(Y_i^0 \neq T_i | X_i) \\ &= pc[1 - H(\beta^t X_i)] + p(1 - c)H(\beta^t X_i), \end{aligned}$$

and

$$\begin{aligned}
P(\delta_i = 0 | X_i) &= P(Y_i^0 = T_i | X_i) \\
&= P(Y_i^0 = 0, T_i = 0 | X_i) + P(Y_i^0 = 1, T_i = 1 | X_i) \\
&= p(1 - c)[1 - H(\beta^T X_i)] + pcH(\beta^T X_i) + (1 - p)(1 - c) + c(1 - p) \\
&= 1 - \{pc[1 - H(\beta^T X_i)] + p(1 - c)H(\beta^T X_i)\}.
\end{aligned}$$

Here the value of Y_i is missing from the cross table of (Y_i^0, T_i) when it cannot be determined as $Y_i^0 = T_i$. Unlike the existing classic RRT, a few Y_i s can be observed, and the missing rate in our situation can be provided by $P(\delta_i = 0 | X_i)$, which depends on X_i , β , p , and c .

In the presence of missing outcome Y_i for model (1), many approaches have been developed for the case where a surrogate of an outcome is observed, such as Pepe (1992), Pepe et al. (1994), Chu and Halloran (2004), and Chen and Breslow (2004). We propose a joint conditional likelihood method that combines the validation and non-validation data to achieve higher efficiency. The method can be viewed as an extension of the methods of Wang et al. (2002), Lee et al. (2012), and Hsieh et al. (2013). With a bit of algebra, it can be shown that the conditional probability of Y_i given $\delta_i = 1$ is

$$\begin{aligned}
P(Y_i = 1 | X_i, \delta_i = 1) &= \frac{P(Y_i = 1, \delta_i = 1 | X_i)}{P(Y_i = 1, \delta_i = 1 | X_i) + P(Y_i = 0, \delta_i = 1 | X_i)} \\
&= H\left\{\beta^T X_i + \ln\left[\frac{1 - c}{c}\right]\right\} \equiv H_+(X_i; \beta).
\end{aligned}$$

Meanwhile, the principle of the non-validation likelihood method is to estimate β by using the probability of Y_i^0 given $\delta_i = 0$. It can be shown that the conditional probability of Y_i^0 given $\delta = 0$ can be expressed as

$$P(Y_i^0 = 1 | X_i, \delta_i = 0) = \frac{pcH(\beta^T X_i) + (1 - p)c}{1 - \{pc[1 - H(\beta^T X_i)] + p(1 - c)H(\beta^T X_i)\}} \equiv H_-(X_i; \beta).$$

Based on $H_+(X_i; \beta)$ and $H_-(X_i; \beta)$, we consider the likelihood function

$$\prod_{i=1}^n H_+(X_i; \beta)^{\delta_i Y_i} [1 - H_+(X_i; \beta)]^{\delta_i (1 - Y_i)} H_-(X_i; \beta)^{(1 - \delta_i) Y_i^0} [1 - H_-(X_i; \beta)]^{(1 - \delta_i)(1 - Y_i^0)}.$$

Thus the score function is as follows:

$$U_{J,n}(\beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \delta_i X_i [Y_i - H_+(X_i; \beta)] + (1 - \delta_i) X_i D_i(\beta) [Y_i^0 - H_-(X_i; \beta)] \right\},$$

where

$$D_i(\beta) = \frac{H^{(1)}(\beta^T \mathcal{X}_i)[pc(1 - pc) + p(1 - p)c(1 - 2c)]}{H_-^{(1)}(X_i; \beta) \{1 - \{pc[1 - H(\beta^T \mathcal{X}_i)] + p(1 - c)H(\beta^T \mathcal{X}_i)\}\}^2}$$

for $H_-^{(1)}(X_i; \beta) = H_-(X_i; \beta)[1 - H_-(X_i; \beta)]$. The joint conditional likelihood estimator of β , denoted by $\hat{\beta}_J$, can be obtained by using the Newton-Raphson method to solve the estimating equations $U_{J,n}(\beta) = \mathbf{0}$. As we before, let $S_i(\beta) = \delta_i \mathcal{X}_i [Y_i - H_+(X_i; \beta)] + (1 - \delta_i) \mathcal{X}_i D_i(\beta) [Y_i^0 - H_-(X_i; \beta)]$. It can then be shown that $\sqrt{n}(\hat{\beta}_J - \beta)$ is asymptotically normally distributed with mean $\mathbf{0}$ and covariance matrix J . Note that when $\pi = 1$ or T_i is observed, the joint conditional likelihood estimator is consistent.

4.2 Design consideration

We begin by reviewing and comparing the joint conditional likelihood method available to researchers. Based on (Y_i^0, X_i, T_i) , $i = 1, 2, \dots, n$, by conducting a cross table, a few responses on the sensitive characteristic Y_i can be elicited from (Y_i^0, T_i) , especially because the sensitive characteristic of individuals is used in the joint conditional likelihood method. Given that, the perceived protection of the respondents can be manipulated. Although the subjective privacy protection is more important than the true statistical privacy protection (Lensvelt-Mulders et al. 2005), in this work, there is no harm in using this information of a few responses on a sensitive nature for analyses, and it is not important to gain access to personal information about the respondents. Leysieffer and Warner (1976) considered the related-question RRT, where if data are collected from direct questions, randomization needs to be carried out a posteriori to protect privacy based on revealing probabilities. From a practical point of view, the respondents' perceptions are significantly correlated with willingness to participate in the research, so it may be questionable whether the respondent is able to measure the extent of privacy disclosure offered by an RRT design. Moreover, some researchers prefer designs privileging a more efficient estimator rather than those guaranteeing higher privacy protection.

Using the approach to collect answers of the unrelated question of RRT of Greenberg et al. (1969) and innocuous question B, we should fully inform the respondents in advance that a few responses on a sensitive characteristic will be elicited from the cross table of answers to the two questions. However, this idea is dangerous because the respondents may feel betrayed by the researchers and many researchers may find that the joint conditional likelihood method is unethical. For protection of the respondents' privacy, we do not recommend deliberately collecting information on the innocuous question directly. To further assuage concerns of the respondents' noncompliance with RRT survey instructions, information collected will be de-identified and statistical only, and will not reveal individual's answer.

5 Simulation study

We conduct a simulation study to evaluate finite-sample performance of the proposed estimator and the maximum likelihood estimator. Hence three estimators are considered as follows:

- $\hat{\beta}_M$: the maximum likelihood (ML) estimator
- $\hat{\beta}_T$: the transformation method (TM) estimator
- $\hat{\beta}_J$: the joint conditional likelihood method (JCLM) estimator

One thousand replications are conducted, and the sample size is $n = 1000$. For each estimator, we compute bias, asymptotic standard error (ASE), standard deviation (SD), and coverage probability of a 95 % confidence interval (CP).

We consider the case with covariates X and Z . First, X is generated from a uniform $[-1, 1]$ distribution, and Z is generated from a binary distribution with $P(Z = 1) = 0.5$. Given X and Z , Y is a binary variable with $P(Y = 1|X, Z) = H(\beta_0 + \beta_1 X + \beta_2 Z)$, where $\beta = (\beta_0, \beta_1, \beta_2)^t = (-\log(2), \log(2), \log(3))^t$, and T is a binary variable with $P(T = 1|X, Z) = c$. Given Y and T , Y^0 is generated as $P(Y^0 = Y) = p$ and $P(Y^0 = T) = 1 - p$. Given T , T^0 is generated from $P(T^0 = T) = \pi$ and $P(T^0 = 1 - T) = 1 - \pi$. Note that the JCLM is valid only when T_i is observed or $\pi = 1$. When $\pi < 1$, we only obtain T_i^0 , and T_i cannot be elicited by T_i^0 . Here if we replace T_i by T_i^0 in the pairs $\{Y_i^0 = 0, T_i^0 = 1\}$ and $\{Y_i^0 = 1, T_i^0 = 0\}$, the inequality between Y_i^0 and T_i^0 can still not provide the valid value of Y_i . If we treat $Y_i = 1$ and $Y_i = 0$ as the pairs $\{Y_i^0 = 1, T_i^0 = 0\}$ and $\{Y_i^0 = 0, T_i^0 = 1\}$, respectively, both the treated values of Y_i can be correct or not. Hence if using the incorrect values of Y_i to plug in the score function $U_{J,n}(\beta)$, the JCLM estimator might be inconsistent and, hence, give misleading results.

As seen in Table 1, we consider $p = 0.4$ or 0.5 , $c = 0.25$ and $\pi = 1$ or 0.9 . Note that p is the probability the respondent choosing the sensitive question by the unrelated-question RRT of Greenberg et al. (1969), c is the probability of a “Yes” answer to the innocuous question, and π is the probability of the respondent choosing the innocuous question by the related-question RRT of Warner (1965). We study the performance of the three estimators under different values of p and π . The simulation results given in Table 1 show that the efficiencies of all the estimators increase as p increases. Note that the ML method does not use the information of T_i^0 , which is independent of π . The TM estimator performs better than the ML estimator. For $\pi = 1$, the TM estimator outperforms the other two estimators. The JCLM estimator performs slightly better than the ML estimator because a few cases of Y can be observed. Here the result of JCLM is jeopardizing privacy. For $\pi = 0.9$, T_i^0 is only observed, and T_i cannot be elicited by T_i^0 . Based on a few cases of Y by treating $Y_i = 1$ and $Y_i = 0$ as the pairs $\{Y_i^0 = 1, T_i^0 = 0\}$ and $\{Y_i^0 = 0, T_i^0 = 1\}$, the Y_i 's are incorrect. The results of Table 1 show the JCLM estimator is inconsistent and misleading results when $\pi = 0.9$.

As seen in Table 2, when $p = 0.5$, $c = 0.25$ or 0.5 , we study the performance of the TM estimator under different values of π . The simulation results show that the efficiencies of all estimators increase as c decreases and π increases. The TM estimator has smaller ASE when $\pi = 0.9$ compared to $\pi = 0.6$. For $\pi < 0.8$, the TM estimator performs slightly better than the ML estimator.

In summary, based on the need for privacy protection, an innocuous question from the unrelated-question RRT was designed by using the related-question RRT of Warner (1965) that provides additional information for developing a more efficient estimation method. The TM estimator performs better than the ML estimator, and the efficiencies increase as π increases. However, when $\pi = 1$, the cross table of (Y_i^0, T_i) provides some useful information about responses to the sensitive question. The JCLM estimator is asymptotically the most efficient compared to the ML estimator. It can be seen that the efficiencies of the TM estimator are strongly influenced by the value of the probability (p), the respondent choosing the sensitive question by the unrelated-question RRT, the probability (c) that she/he answers “Yes” to the innocuous question, and π that is the probability of the respondent choosing the innocuous question by the related-question RRT of Warner (1965). Note that the three probabilities p , c and π are assumed to be known and can be handled in the survey design. The c was obtained from available records, on a group basis, before the survey was administered.

6 Example

In this section, we describe how to analyze the extramarital relationship study data and cable TV study data. It is straightforward to see that we could view a demographic question to indirectly elicit information on the innocuous question from the unrelated-question RRT after the randomized response data are collected. However, we want to fill this gap by providing a suite of our methodological tools that facilitate the use of RRT in applied research. Therefore, based on the fact that T_i is completely observed, we will compare the performance of the various probabilities of the respondents choosing the innocuous question by conducting simulation experiments. From the previous section, we learn that the TM estimator is asymptotically the most efficient, and the JCLM estimator offers an ideal model with $\pi = 1$.

6.1 Extramarital relationship data

We wanted to estimate the proportion of people in extramarital relationships in Taiwan. The data were collected as part of the TSCS from Year 6 of Cycle 3 in 2012 that includes the unrelated-question RRT design concerning the experience of extramarital relationships. The two questions were designed as follows:

A: Have you ever had sex with someone other than your spouse/partner?

B: Were you born in the month of January, February, or March?

Please pick up one card from the deck of playing cards and do not tell the interviewer the number on the playing card. Remember the card then please answer the question A or B according to the number on the playing card. If the number on the playing card is 1, 2, or 3,

please answer question A. If the number on the playing card is 4 or 5, please answer question B. Note that the probability of answering the sensitive question A is $p = 0.5$. In this survey, there is a demographic question as follows:

C: When were you born?

Because the question C is related to the innocuous question B, it could be used to elicit T_i and reflects $c = 0.25$. In particular, because T_i is completely observed, we apply the classic design by Warner (1965) to simulate T_i^0 from T_i with probability π and $1 - T_i$ with probability $1 - \pi$, i.e., $T_i^0 = W_i T_i + (1 - W_i)(1 - T_i)$ and $W_i \sim \text{Bernoulli}(1, \pi)$. The benchmark results related $\pi = 0.9$ and 0.8 are the average of results generated from 10 replications, and in each replication we conduct simulation experiments to generate T_i^0 from the above procedures. These results are presented in the last two columns of Table 3.

We conduct a logistic regression analysis by using the dichotomous variables, gender and attitude toward extramarital sex as follows:

$$P(Y_i = 1 | X_i, Z_i) = H(\beta_0 + \beta_1 X_i + \beta_2 Z_i), i = 1, 2, \dots, 1847.$$

Note that the covariate X is used to denote gender (1 = male; 0 = female), and Z is used to denote extramarital sex attitudes as follows: “Do you think a married man (woman) may have sex with someone other than his wife (her husband)?” (1 = Yes, 0 = No). The analysis is aimed to estimate the parameter vector $\beta (\beta_0, \beta_1, \beta_2)^t$. There are 1,847 subjects; among them 425 subjects are in the validation data set from the cross table of (Y^0, T) , in which the missing rate of Y is 77 %.

The results are given in Table 3, where $\hat{\beta}_M$, $\hat{\beta}_J$, and $\hat{\beta}_T$ designate the ML estimator, the JCLM estimator, and the TM estimator, respectively. For $\pi = 1$, all the estimates of β_1 are significant, and the estimated ASE of the TM estimator is the smallest. The estimate of β_2 , obtained based on the JCLM, is not significant. The results of the ML method and TM show the significant effects of gender and extramarital sex attitudes on the proportion of extramarital relationships. Moreover, we study the performance of the TM estimator as $\pi = 0.8$ and 0.9 . The estimated ASE of the TM estimator increases as π decreases because it is a protected attribute. Finally, as a benchmark we present a computer simulation result that has two cases related to $\pi = 0.9$ and 0.8 on 10 replications.

6.2 Cable TV data

We studied the proportion of unauthorized cable TV users in Taichung, Yunlin, and Taipei, Taiwan in 2004, from Hsieh et al. (2009). Because unauthorized cable TV use is illegal and a sensitive question, the unrelated-question RRT design was used in this telephone interview survey. Respondents were required to answer one of the following two questions based on the last digit of their identification number. We asked the following two questions:

A: Is your cable TV connection unauthorized?

B: Is the number of family members living with you an odd number?

If the last digit of the respondent's identification number was an odd number, she/he had to answer question A; otherwise she/he answered question B. Here only the respondent knew whether the last digit of her/his identification number was odd. The probability of answering question A is equal to $p = 0.5$. Moreover, we could view a demographic question in the same survey as follows:

C: Including you, how many people live in your house?

The placement of question C could have an effect on responses. The number of people living in your house (including you) is supposedly observable. Therefore, question C related to question B provides valuable information that could be used to elicit T and reflects $c = 0.45$. A random sample of 1482 subjects were independently selected from houses or apartments that needed a cable TV connection. There are 516 subjects in the validation data set in which the missing rate of Y is 65 %. The covariate X is used to denote cities in Taiwan (1. Taichung; 2. Yunlin; 3. Taipei). We define the dummy variable vector (DX_1, DX_2) for city of residence X in which (1, 0) is Taichung, (0, 1) is Yunlin, and (0, 0) is Taipei. We consider the following logistic regression model:

$$P(Y_i = 1 | X_i) = H(\beta_0 + \beta_1 DX_{1i} + \beta_2 DX_{2i}), i = 1, 2, \dots, 1482.$$

The analysis is designed to estimate the parameter vector $\beta = (\beta_0, \beta_1, \beta_2)'$. The results are given in Table 4; for $\pi = 1$, all the estimates of β_1 and β_2 based on all the methods are significant except the ML estimator. This means that the proportion of unauthorized cable TV users in Taipei is different from those in Taichung and Yunlin. The TM and JCLM estimates are very close, and the estimated ASE of the JCLM estimator is smaller than the others. As before, based on the fact that T_i is completely observed, we conduct a simulation study to compare the performance. Note that T_i^0 is generated from T_i with probability π and $1 - T_i$ with probability $1 - \pi$. The simulation results for various values of π show –that the estimated ASE of– the TM estimator increases as π decreases, which is consistent with the simulation results in Sect. 5. Finally, the benchmark results related to $\pi = 0.9$ and 0.8 are the average of results generated from 10 replications, and in each replication we conduct simulation experiments to generate T_i^0 from the above procedures. These results are presented in the last two columns of Table 4.

In these two examples, we control $p = 0.5$ and let c vary from an extreme value (the extramarital relationship example: $c = 0.25$) to the middle range (the cable TV example: $c = 0.45$). The two real data examples show that the TM estimator is more efficient than the ML estimator when $\pi = 1$. Because T_i is completely observed, the simulation results show that the estimated ASE of the TM estimator increases as π decreases because it is a protected attribute. Therefore, the results of the two real data examples are consistent with those of the simulation study, and the innocuous question with c known was appropriate for these data. If one has no idea whatsoever on which side of 0.5 to expect the probability that she/he answers “Yes” to the sensitive question, then a moderate value of c between 0.25 and 0.75 will at least control the loss in efficiency; see Greenberg et al. (1969).

7 Conclusions

In this work, we considered the joint use of the unrelated-question RRT of Greenberg et al. (1969) and the related-question RRT of Warner (1965) dealing with the issue of an innocuous question from the unrelated-question RRT. The approach can provide additional information on an innocuous question for developing a more efficient estimation method of a logistic regression model for the binary outcome variable, a sensitive characteristic. The additional information can effectively improve the efficiency of the maximum likelihood estimator of Scheers and Dayton (1988). From a practical point of view, the subjective privacy protection is more important than the true statistical privacy protection (Lensvelt-Mulders et al. 2005). In this new design, information collected will be de-identified and statistical only, and will not reveal individual's answer. Additionally, to increase the efficiency of privacy protection, the new design involves fully informing the respondents in advance about the consequences of the survey in the future.

However, for the case of the two real data examples, under the assumption of all truthful response by respondents, we could view a demographic question to indirectly elicit information on the innocuous question from the unrelated-question RRT after the randomized response data are collected. Hence the additional information is different from the related-question RRT of Warner (1965) dealing with the issue of an innocuous question from the unrelated-question RRT. Moreover, from a practical point of view, the basic demographic question was always asked first, and the unrelated question of RRT was asked last to ensure an acceptable response rate in practice. When the respondent is willing to answer the unrelated question of RRT, she/he might be more willing to cooperate with the interviewer. In other words, she/he believes that her/his privacy is protected or her/his perceived risk of disclosure is low.

In this work, the usefulness of information provided by respondents must be determined on the level of participation of the respondents' psychology, the costs of surveying and other factors. To improve its efficiency, we use the probability (p) that the respondent chooses the sensitive question, the probability (c) that she/he answers "Yes" to the innocuous question, and the probability (π) of the respondent choosing the innocuous question. To compare the performance among the methods, we present the transformation method and joint conditional likelihood method. When the answer to the innocuous question can be completely observed, the joint conditional likelihood method depends on the loss of *a posteriori* privacy from the cross table because in this method a few cases are indirectly elicited to provide answers to the sensitive question. When the answer to the innocuous question cannot be completely observed, the joint conditional likelihood method is inconsistent and gives misleading results. The results show that the estimated ASE of the transformation method estimator increases as π decreases because it is a protected attribute. The simulation results demonstrate that in terms of efficiency the transformation method outperforms the maximum likelihood method when $\pi = 0.8$. Those measures are mathematical comparisons of efficiency. The possible extension of the proposed method would be a logistic regression model when the covariates are missing.

Acknowledgments

The authors are grateful to an Associate Editor and a referee for their helpful comments that improved the presentation. The research of S.H. Hsieh and S.M. Lee was supported by the National Science Council (NSC) of Taiwan, ROC (100–2118-M-001–001-MY2 and 101–2118-M-035–004-MY2, respectively). The project described was also supported by NSC Taiwan, ROC, through grant 100–2420-H-001–017-MY2 (S.H. Hsieh and S.M. Lee), and by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), through Grant #UL1 TR000002, and MIND Institute Intellectual and Developmental Disabilities Research Center (U54 Hd079125) (C.S. Li). The authors would like to thank research fellows Ying-Hwa Chang for adding the RRT questionnaire design by the research project “Taiwan Social Change Survey: Year 6 of cycle 3”, which was sponsored by the NSC. The Survey Research Data Archive, Academia Sinica is responsible for the data distribution. The authors appreciate the assistance of the aforementioned institutes and individuals in providing all relevant data.

References

- Chang H, Wang C, Haung K (2004) On estimating the proportion of a qualitative sensitive character using randomized response sampling. *Qual Quant* 38:675–680
- Chen J, Breslow NE (2004) Semiparametric efficient estimation for the auxiliary outcome problem with the conditional mean model. *Can J Stat* 32:359–372
- Chu H, Halloran ME (2004) Estimating vaccine efficacy using auxiliary outcome data and a small validation sample. *Stat Med* 23:2697–2711 [PubMed: 15316953]
- Corstange D (2004) Sensitive questions, truthful response? Randomized response and hidden logit as a procedure to estimate it In: Annual meeting of the American Political Science Association. Chicago, 2–5 9 2004, <http://www.umich.edu/dancorst/>
- Foutz RV (1977) On the unique consistent solution to the likelihood equations. *J Am Stat Assoc* 72:147–148
- González BJ, von Davier M (2013) Statistical models and inference for the true equating transformation in the context of local equating. *J Educ Meas* 50(3):315–320
- Gribble JN, Miller HG, Rogers SM, Turner CF (1999) Interview mode and measurement of sexual behaviors: methodological issues. *J Sex Res* 36(1):16–24 [PubMed: 23226876]
- Gjestvang CR, Singh S (2006) A new randomized response model. *J R Stat Soc Ser B* 68:523–530
- Greenberg BG, Abul-Ela A, Simmons WR, Horvitz DG (1969) The unrelated question randomized response model: theoretical framework. *J Am Stat Assoc* 64:520–539
- Haung K (2004) A survey technique for estimating the proportion and sensitivity in a dichotomous finite population. *Stat Neerlandica* 58:75–82
- Horvitz DG, Shah BV, and Simmons WR (1967) The unrelated question randomised response model. In: Proceedings of the social statistics section, American Statistical Association, pp 65–72
- Hsieh SH, Lee SM, Shen PS (2009) Semiparametric analysis of randomized response data with missing covariates in logistic regression. *Comput Stat Data Anal* 53:2673–2692
- Hsieh SH, Li CS, Lee SM (2013) Logistic regression with outcome and covariates missing separately or simultaneously. *Comput Stat Data Anal* 66:32–54
- Kim JM, Heo TY (2013) Randomized response group testing model. *J Stat Theory Pract* 7:33–48
- Kim JM, Warde WD (2004) A stratified Warner’s randomized response model. *J Stat Plann Inference* 120:155–165
- Kuk AYC (1990) Asking sensitive questions indirectly. *Biometrika* 77:436–438
- Lee SM, Li CS, Hsieh SH, Huang LH (2012) Semiparametric estimation of logistic regression model with missing covariate and outcome. *Metrika* 75:621–653
- Lensvelt-Mulders GJLM, Hox JJ, van der Heijden PGM, Maas CJM (2005) Meta-analysis of randomized response research: thirty-five years of validation. *Sociol Methods Res* 33:319–334
- Leyseffer FW, Warner SL (1976) Respondent jeopardy and optimal designs in randomized response models. *J Am Stat Assoc* 71:649–656
- Mangat NS, Singh R (1990) An alternative randomized response procedure. *Biometrika* 77:439–442
- Magder LS, Hughes JP (1997) Logistic regression when the outcome is measured with uncertainty. *J Am Stat Assoc* 146:195–203

- Mangat NS (1994) An improved randomized response strategy. *J R Stat Soc Ser B* 56:93–95
- Moors JJA (1971) Optimization of the unrelated question randomized response model. *J Am Stat Assoc* 66:627–629
- Pepe MS (1992) Inference using surrogate outcome data and a validation sample. *Biometrika* 79:355–365
- Pepe MS, Reilly M, Fleming TR (1994) Auxiliary outcome data and the mean-score method. *J Stat Plan Inference* 42:137–160
- Raghavarao D (1978) On an estimation problem in Warner’s randomized response technique. *Biometrics* 34:87–90
- Scheers NJ, Dayton CM (1988) Covariate randomized response models. *J Am Stat Assoc* 83:969–974
- Singh S, Singh R, Mangat NS (2000) Some alternative strategies to Moor’s model in randomized response sampling. *J Stat Plan Inference* 83:243–255
- Van den Hout A, Van der Heijden PGM, Gilchrist R (2007) The logistic regression model with response variables subject to randomized response. *Comput Stat Data Anal* 51:6060–6069
- Wang CY, Chen JC, Lee SM, Ou ST (2002) Joint conditional likelihood estimator in logistic regression with missing covariate data. *Stat Sin* 12:555–574
- Warner SL (1965) Randomized response: a survey technique for eliminating evasive answer bias. *J Am Stat Assoc* 60:63–69 [PubMed: 12261830]

Fig. 1.

Probability mass diagrams of answers from the unrelated-question RRT of Greenberg et al. (1969) and the related-question RRT of Warner (1965) dealing with the issue of an innocuous question from the unrelated-question RRT

Table 1Simulation results for various values of π and p with $c = 0.25$

π	Parameter		$p = 0.4$			$p = 0.5$		
			$\hat{\beta}_M$	$\hat{\beta}_J$	$\hat{\beta}_T$	$\hat{\beta}_M$	$\hat{\beta}_J$	$\hat{\beta}_T$
1	β_0	Bias	-0.0254	0.0046	-0.0122	-0.0179	0.0027	-0.0118
		SD	0.2430	0.2148	0.1910	0.1924	0.1853	0.1629
		ASE	0.2432	0.2131	0.1925	0.1929	0.1853	0.1661
		CP	0.9530	0.9450	0.9460	0.9550	0.9550	0.9570
	β_1	Bias	0.0273	0.0173	0.0172	0.0189	0.0136	0.0134
		SD	0.3115	0.2777	0.2646	0.2488	0.2444	0.2259
		ASE	0.3108	0.2749	0.2579	0.2452	0.2372	0.2186
		CP	0.9550	0.9380	0.9510	0.9550	0.9430	0.9460
	β_2	Bias	0.0445	0.0205	0.0308	0.0246	0.0083	0.0187
		SD	0.3516	0.3121	0.2993	0.2767	0.2689	0.2485
		ASE	0.3489	0.3126	0.2928	0.2762	0.2700	0.2480
		CP	0.9560	0.9450	0.9390	0.9570	0.9450	0.9510
0.9	β_0	Bias	-0.0254	-0.4036	-0.0166	-0.0179	-0.3250	-0.0148
		SD	0.2430	0.1901	0.2057	0.1924	0.1752	0.1693
		ASE	0.2432	0.1890	0.2081	0.1929	0.1745	0.1746
		CP	0.9530	0.4210	0.9440	0.9550	0.5410	0.9550
	β_1	Bias	0.0273	0.0459	0.0197	0.0189	0.0508	0.0145
		SD	0.3115	0.2443	0.2810	0.2488	0.2265	0.2344
		ASE	0.3108	0.2397	0.2746	0.2452	0.2204	0.2273
		CP	0.9550	0.9450	0.9480	0.9550	0.9390	0.9520
	β_2	Bias	0.0445	0.0770	0.0346	0.0246	0.0774	0.0214
		SD	0.3516	0.2718	0.3136	0.2767	0.2486	0.2543
		ASE	0.3489	0.2726	0.3102	0.2762	0.2511	0.2570
		CP	0.9560	0.9440	0.9490	0.9570	0.9350	0.9570

The true value of $\beta = (-\log(2), \log(2), \log(3))$ T and $n = 1000$

Table 2Simulation results for various values of π with $p = 0.5$

c	Parameter		$\hat{\beta}_M$	$\pi = 1$	$\pi = 0.9$	$\pi = 0.8$	$\pi = 0.7$	$\pi = 0.6$
				$\hat{\beta}_T$	$\hat{\beta}_T$	$\hat{\beta}_T$	$\hat{\beta}_T$	$\hat{\beta}_T$
0.25	β_0	Bias	-0.0179	-0.0118	-0.0148	-0.0159	-0.0184	-0.0183
		SD	0.1924	0.1629	0.1693	0.1783	0.1844	0.1890
		ASE	0.1929	0.1661	0.1746	0.1820	0.1879	0.1916
		CP	0.9550	0.9570	0.9550	0.9530	0.9540	0.9560
	β_1	Bias	0.0189	0.0134	0.0145	0.0170	0.0174	0.0193
		SD	0.2488	0.2259	0.2344	0.2396	0.2420	0.2461
		ASE	0.2452	0.2186	0.2273	0.2348	0.2403	0.2440
		CP	0.9550	0.9460	0.9520	0.9530	0.9530	0.9580
	β_2	Bias	0.0246	0.0187	0.0214	0.0236	0.0258	0.0256
		SD	0.2767	0.2485	0.2543	0.2639	0.2690	0.2737
		ASE	0.2762	0.2480	0.2570	0.2649	0.2709	0.2749
		CP	0.9570	0.9510	0.9570	0.9570	0.9560	0.9540
0.5	β_0	Bias	-0.0196	-0.0138	-0.0163	-0.0177	-0.0203	-0.0197
		SD	0.2104	0.1755	0.1890	0.1988	0.2051	0.2087
		ASE	0.2106	0.1787	0.1921	0.2009	0.2066	0.2096
		CP	0.9540	0.9520	0.9460	0.9470	0.9510	0.9550
	β_1	Bias	0.0168	0.0125	0.0133	0.0166	0.0172	0.0173
		SD	0.2699	0.2370	0.2480	0.2596	0.2656	0.2684
		ASE	0.2573	0.2279	0.2404	0.2486	0.2537	0.2564
		CP	0.9430	0.9430	0.9350	0.9360	0.9370	0.9410
	β_2	Bias	0.0318	0.0254	0.0280	0.0306	0.0335	0.0321
		SD	0.2898	0.2571	0.2696	0.2797	0.2842	0.2884
		ASE	0.2902	0.2580	0.2714	0.2804	0.2862	0.2893
		CP	0.9560	0.9480	0.9480	0.9520	0.9530	0.9530

The true value of $\beta = (-\log(2), \log(2), \log(3))$ T and $n = 1000$

Table 3Analysis results of extramarital relationship data for various values of π

Variable	Parameter	$\hat{\beta}_M$	$\pi = 1$		$\pi = 0.9$		$\pi = 0.8$		Benchmark	
									$\pi = 0.9$	$\pi = 0.8$
			$\hat{\beta}_J$	$\hat{\beta}_T$	$\hat{\beta}_T$	$\hat{\beta}_T$	$\hat{\beta}_T$	$\hat{\beta}_T$	$\hat{\beta}_T$	$\hat{\beta}_T$
Intercept	β_0	-1.6043* (0.1921)	-1.6399* (0.1491)	-1.4573* (0.1386)	-1.4994* (0.1527)	-1.6320* (0.1750)	-1.4989* (0.1550)	-1.5361* (0.1684)		
X	β_1	0.9978* (0.2289)	0.9835* (0.1882)	0.9627* (0.1770)	0.9468* (0.1912)	1.0413* (0.2111)	0.9684* (0.1929)	0.9614* (0.2057)		
Z	β_2	0.8239* (0.3185)	0.5166* (0.2805)	0.7868* (0.2863)	0.8361* (0.2936)	0.9567* (0.3070)	0.8184* (0.2983)	0.8348* (0.3055)		

$n = 1847$, $p = 0.5$ and $c = 0.25$. “*” denotes significant estimates of the parameters. The benchmark results are generated from multiple imputation 10 times. Values in parentheses () are the asymptotic standard error of the estimates. X is the variable for gender (1 = Male; 0 = Female); Z is the variable for extramarital sex attitudes (1 = Yes; 0 = No)

Table 4Analysis results of cable TV data for various values of π

Variable	Parameter	$\hat{\beta}_M$	$\pi = 1$		$\pi = 0.9$		$\pi = 0.8$		Benchmark	
									$\pi = 0.9$	$\pi = 0.8$
			$\hat{\beta}_J$	$\hat{\beta}_T$	$\hat{\beta}_T$	$\hat{\beta}_T$	$\hat{\beta}_T$	$\hat{\beta}_T$	$\hat{\beta}_T$	$\hat{\beta}_T$
Intercept	β_0	-2.8719* (0.6103)	-1.8173* (0.1523)	-1.6395* (0.1553)	-1.9278* (0.2167)	-2.2594* (0.3116)	-1.9118* (0.2153)	-2.2419* (0.3085)		
DX_1	β_1	1.1089 (0.7682)	0.5784* (0.2722)	0.5683* (0.2812)	0.7385* (0.3473)	0.7736 (0.4670)	0.6657 (0.3565)	0.7955 (0.4596)		
DX_2	β_2	1.0229 (0.7126)	0.6971* (0.2178)	0.7083* (0.2257)	0.7709* (0.2928)	0.8274* (0.3978)	0.7351* (0.2942)	0.8299* (0.3950)		

Note: $n = 1482$, $p = 0.5$ and $c = 0.45$. “*” denotes significant estimates of the parameters. The benchmark results are generated from multiple imputation 10 times. Values in parentheses () are the asymptotic standard error of the estimates. DX_1 and DX_2 are dummy variables for city of residence X (1 = Taichung; 2 = Yunlin; 3 = Taipei)