mathematics

MDPI

Review

Randomized Response Techniques: A Systematic Review from the Pioneering Work of Warner (1965) to the Present

Truong-Nhat Le [1], Shen-Ming Lee [2], Phuoc-Loc Tran [3] and Chin-Shang Li [4,*]

[1] Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam; letruongnhat@tdtu.edu.vn
[2] Department of Statistics, Feng Chia University, Taichung 40724, Taiwan; smlee@mail.fcu.edu.tw
[3] Department of Mathematics, College of Natural Science, Can Tho University, Can Tho 900000, Vietnam; tploc@ctu.edu.vn
[4] School of Nursing, The State University of New York, Buffalo, NY 14214, USA
* Correspondence: csli2003@gmail.com or chinshan@buffalo.edu

Abstract: The randomized response technique is one of the most commonly used indirect questioning methods to collect data on sensitive characteristics in survey research covering a wide variety of statistical applications including, e.g., behavioral science, socio-economic, psychological, epidemiology, biomedical, and public health research disciplines. After nearly six decades since the technique was invented, many improvements of the randomized response techniques have appeared in the literature. This work provides several different aspects of improvements of the original randomized response work of Warner, as well as statistical methods used in the RR problems.

Keywords: indirect questioning; non-randomized response technique; randomized response technique; sensitive attribute; statistical methods

MSC: 62F12; 62F15; 62J12

Citation: Le, T.-N.; Lee, S.-M.; Tran, P.-L.; Li, C.-S. Randomized Response Techniques: A Systematic Review from the Pioneering Work of Warner (1965) to the Present. *Mathematics* **2023**, *11*, 1718. https://doi.org/10.3390/math11071718

Academic Editor: Christophe Chesneau

Received: 12 January 2023
Revised: 24 March 2023
Accepted: 1 April 2023
Published: 3 April 2023

1. Introduction to Randomized Response Techniques

Sample surveys are commonly used to collect data for studies in a wide range of statistical applications such as behavioral science, socio-economic, psychological, epidemiological, biomedical, and public health research disciplines. Mail surveys, telephone surveys, and personal interviews (face-to-face interviews) are the commonly used traditional data-collection methods; see, e.g., [1]. The data collected from these surveys are used to estimate and make statistical inferences about the unknown population parameters of interest, e.g., the population proportion of individuals with a certain property appearing in most related research, the population honest response rate [2,3], and the sensitivity level of a question of interest; in other words, the population proportion of individuals considering the question of interest to be sensitive [4–6]. Because of that, researchers and practitioners are particularly interested in the reliability of collected data (e.g., non-response rate and dishonest answer rate) in studies using sample surveys, but more so while the topics of investigation involve, e.g., threatening, embarrassing, stigmatizing, highly personal, and even incriminating issues. The aforementioned issues are collectively referred to as sensitive characteristics (attributes, behaviors, features, traits). For example, people consider abortion behavior, cheating on examinations, discrimination, domestic violence, drug use, gambling, illegal income, plagiarism, political opinions, sexual behavior, tax evasion, and other illicit behaviors to be sensitive. Refer to [7] for a more detailed classification of the three types of sensitive questions.

Research on sensitive issues is increasingly receiving the attention of many researchers, practitioners, and social organizations. For instance, in the study by Krumpal and Voss [8],

the General Social Survey (Allgemeine Beölkerungsumfrage der Sozialwissenschaften—ALLBUS) in Germany asked respondents whether they have committed tax evasion and shoplifting, dodged fares, or driven drunk. In the United States, the National Survey on Drug Use and Health (NSDUH) and the General Social Survey (GSS) routinely require surveyees to self-report on sensitive issues, e.g., sexual habits or drug use. The Taiwan Social Change Survey (TSCS) conducted face-to-face interviews about sexual orientation [9,10], the presidential election [11], monthly income [12], and extramarital relationships [13,14]. Estimating the prevalence of such sensitive features is of great importance in helping researchers to build scientific knowledge and recommend necessary strategies to the authorities.

It is widely accepted that most survey participants consider the aforementioned issues to be secret, shameful, and even illegal. Then, when participating in surveys that use traditional data-collection methods including, e.g., computer-assisted self-interviewing or telephone interviewing and self-administered questionnaires with paper and pencil, to avoid being stigmatized by society or punished by the government, and to leave a good impression on others, survey respondents tend to ignore sensitive questions, which causes a non-response bias problem, or they answer these sensitive questions according to socially desirable behaviors and attitudes, which causes a social desirability bias problem. See, e.g., [7,15,16]. For example, a student is directly asked a question about a socially undesirable behavior: "Have you ever cheated on examinations?". Naturally, regardless of whether she/he has ever cheated on examinations or not, it is more likely she/he may deny it. Refer to [17] for more information on this sensitive topic. Or, in a validation study by Preisendörfer and Wolter [18], where the researchers knew the true answers in advance, 42 percent (face-to-face interviews) and 33 percent (mail survey) of respondents did not admit that they had been convicted. Likewise, van der Heijden et al. [19] conducted a face-to-face interview, and 75 percent of respondents who committed welfare or unemployment benefits fraud denied doing so. As another real example, Hsieh and Perri [20] pointed out that the proportion of non-heterosexual subjects present in a community is generally underestimated if respondents have to answer sensitive questions directly. In contrast, respondents tend to present themselves positively by displaying behaviors and attitudes that conform to social norms, such as engaging in charitable activities, volunteering, and eating healthily. See, e.g., [1]. In general, socially desirable attributes are over-reported while socially undesirable attributes are under-reported when data are collected by direct interrogation methods. Therefore, the quality of data collected through direct questions on such topics is not guaranteed. As a result, collected data may produce inaccurate estimated results and invalid inferences about the sensitive behavior under investigation.

In an effort to reduce potential bias due to social desirability response and non-response and thereby improve the reliability of data gathered from responses to sensitive questions for better estimation of the population proportion of individuals who have a sensitive characteristic, various indirect questioning techniques (IQTs) have been proposed by, e.g., [21–24]. Among them, some commonly used techniques are the randomized response (RR) technique (RRT) [25], the unmatched count technique—also called the item count technique—unmatched block design, or block total response [26,27], and the triangular model (TRM) and crosswise model (CWM) [28], which are two of the non-randomized response (NRR) techniques (NRRTs). These techniques have been developed to ensure anonymity as well as minimize the feelings of jeopardy for survey respondents when answering sensitive questions. That is what motivates them to answer honestly sensitive questions.

Blair et al. [29] provided an excellent review of the RRTs and classified them into mirrored question, forced response, disguised response, and unrelated question techniques. Among these techniques, Sungkawichai et al. [30] extended the classical forced RRT by using an arbitrary random variable. Tian and Tang [31] also presented another classification for the RRTs. Interested readers may also refer to the monographs on the RRTs and other alternative IQTs by [21,32–37] for comprehensive reviews. Tian and Tang [31] contributed

an excellent monograph to the NRRT until 2013. Next, we present a review of RRTs from the work of Warner [25] to the present.

2. Warner's Randomized Response Design and Some Direct Extensions

The first version of the RRT, conceived by Warner [25] in 1965, is to increase the response rate and eliminate dishonest responses for the estimation of the proportion of individuals in a population bearing some sensitive attribute. The main idea of the RRT is to add random noise to respondents' answers for the protection of their privacy. Specifically, according to the idea of Warner [25], two questions were designed: a sensitive question of interest and its complementary question. That is why the original design of [25] is also known as a "related-question RR design". For example,

A : Have you ever had a one-night stand through a dating website or mobile app (with probability p of selecting this question).

$\overline{\mathbf{A}}$: Have you never had a one-night stand through a dating website or mobile app (with probability $1-p$ of selecting this question).

Suppose we wish to estimate the proportion θ of people belonging to a sensitive group, called group A. A simple random sample of size n is selected from the population. Each surveyee uses the outcome generated by a randomization device, e.g., spinners, dice, or random number generators, which is not observed by the interviewer, to determine which question to honestly answer "Yes" or "No" to. The interviewee responds to statements **A** and $\overline{\mathbf{A}}$ with probabilities p and $1-p$, respectively. Let n_1 be the number of individuals responding "Yes". The parameter θ is estimated based on the indirect responses of all individuals via the maximum likelihood (ML) estimator $\widehat{\theta}_W = \frac{p-1}{2p-1} + \frac{n_1}{n(2p-1)}$ with $\mathrm{Var}(\widehat{\theta}_W) = \frac{\theta(1-\theta)}{n} + \frac{p(1-p)}{n(2p-1)^2}$ as long as $p \neq 0.5$. $\widehat{\theta}_W$ is then an unbiased estimator of θ and used to replace θ to obtain an estimator of $\mathrm{Var}(\widehat{\theta}_W)$. See Appendix A.1. Because the surveyor does not know which question has been answered by the interviewee, the respondent can feel more comfortable with answering sensitive questions without fear of personal privacy being revealed. It makes the respondent more likely to give an honest response to the sensitive question in case she/he carries that sensitive characteristic. In fact, a validation study by Lensvelt-Mulders et al. [38] showed that, for sensitive questions, the RRT yields a more valid estimation of prevalence in comparison to other methods.

Despite solving many of the problems posed earlier, the original RR design of Warner [25] has certain limitations. For example, Warner's model does not work for $p = 0.5$. However, the inefficiency of Warner's model is its most serious limitation when compared with the design of *direct questioning* (DQ), which is clearly demonstrated in Tian and Tang [31]. The variance of the estimator $\widehat{\theta}$ of θ by the DQ design is $\mathrm{Var}(\widehat{\theta}) = \frac{\theta(1-\theta)}{n}$, based on the binomial distribution with parameters n and θ. Using the RR design of Warner [25] induces the extra variance, $\frac{p(1-p)}{n(2p-1)^2}$, which is the variance due to the randomization device, compared to $\mathrm{Var}(\widehat{\theta})$. Accordingly, $\widehat{\theta}_W$ is less efficient than $\widehat{\theta}$. During nearly six decades of efforts to overcome these limitations and improve computational efficiency, quite a few alternative RR models have been proposed and empirically applied. For instance, just to name a few, Horvitz et al. [39] and Greenberg et al. [40] combined a sensitive question of interest and another question that is innocuous and completely unrelated to the sensitive topic to propose an unrelated-question RR design. Chaudhuri and Mukerjee [41] introduced optional RRTs. Bhargava and Singh [42] introduced a modified randomization device for the RR design of [25]. Kim and Warde [43] proposed a stratified RR design of [25].

Abbasi et al. [44] proposed a partial RRT to gather reliable sensitive data for the estimation of the proportion of a population in a ranked set sampling scheme using auxiliary information. The authors provided respondents the option of both "direct" and "randomized" responses for the sensitive question in order to increase their confidence/co-operation. Zapata et al. [45] proposed an electronic randomization device, which is able to directly produce a response when utilized by a respondent. The proposed randomization device

builds upon the model of Warner [25] by utilizing a variation on the spinner approach. However, instead of a physical spinner, they have developed a model, which utilizes the Python programming language to electronically replicate the functionality of a spinner with the selection of a button, with the user simply being requested to choose either a "Red" or "Green" button depending on his/her status of possessing a sensitive characteristic.

3. Some Aspects Extended from Warner's Randomized Response Design

3.1. Unrelated-Question Randomized Response Design

Motivated by the case where the model of Warner [25] does not work when $p = 0.5$, Horvitz et al. [39] and Greenberg et al. [40] modified Warner's method by incorporating a non-sensitive question within a sensitive question. Along with that, some respondents find the questions in the design of [25] sensitive or uncomfortable to answer even if a randomization device is used. Two questions in the unrelated-question design, for example, are given as follows:

A : Have you ever had a one-night stand through a dating website or mobile app (with probability p of selecting this question).

C : Were you born between January and September (with probability $1 - p$ of selecting this question).

Again, each respondent selected in the sample uses a device such as a deck of cards to determine the question to which she/he responds. Let c_0 be the true proportion of individuals with non-sensitive characteristic. If c_0 is known, [39,40] proposed the unbiased estimator $\widehat{\theta}_{U_1} = \frac{n_1/n-(1-p)c_0}{p}$ for θ. In the case where c_0 is unknown, they considered two independent samples of sizes n_1^* and n_2^* with $n = n_1^* + n_2^*$. In each sample, the above procedure is carried out. Assume that the probabilities of selecting the designed sensitive question in the samples of sizes n_1^* and n_2^* are p_1 and p_2, respectively, with $p_1 \neq p_2$. They proposed the unbiased estimator $\widehat{\theta}_{U_2} = \frac{(1-p_2)m_1/n_1^*-(1-p_1)m_2/n_2^*}{p_1-p_2}$ for θ, where m_1 and m_2 are the numbers of respondents who answer "Yes" in the first and second samples, respectively. Because the modified method boosts the degree of privacy, it may receive greater cooperation from respondents. According to Edgell et al. [46], compared to the RRT of Warner [25], the unrelated-question RRT is much more statistically efficient and becomes even more so when the population parameters of the non-sensitive question are known. To assess whether respondents would honestly respond to the non-sensitive question, even if it could be interpreted as socially undesirable when paired with a sensitive question, the researchers conducted a study using an unrelated-question RRT. Shaw and Chaudhuri [47] utilized the approach of the inverse hypergeometric trial to improve the revised unrelated characteristics model device of Chaudhuri and Shaw [48]. Lee et al. [14] introduced a data-collection method for survey on sensitive issues in which both the unrelated-question RRT and the DQ design are combined. They proposed two new methods for estimating the proportion of respondents possessing the sensitive attribute under a missing data setup.

3.2. Some Kind of Two-Stage Randomized Response Design

In 1990, Mangat and Singh [49] proposed a two-stage RR procedure in which two randomization devices, R_1 and R_2, are used. In the first stage, each survey participant is asked to use the randomization device R_1, such as a well-shuffled deck of cards, to select one from the following two statements:

A : I belong to group A.

C : go to the randomization device R_2.

The two statements are selected with probabilities p_0 and $1 - p_0$, respectively. In the second stage, the design of Warner [25] is used. An unbiased estimator of θ is shown as $\widehat{\theta}_{MS} = \frac{n_1/n-(1-p_0)(1-p)}{2p-1+2p_0(1-p)}$ with $\mathrm{Var}(\widehat{\theta}_{MS}) = \frac{\theta(1-\theta)}{n} + \frac{(1-p)(1-p_0)\{1-(1-p)(1-p_0)\}}{n\{2p-1+2p_0(1-p)\}^2}$. It is shown that compared to the RR design of Warner [25], the two-stage RR design is more efficient.

Mangat [50] proposed another RR model in which each interviewee is asked to respond "Yes" if she/he were in the sensitive group; she/he is guided to utilize the device of Warner [25] otherwise. It is shown that the RR design of Mangat [50] is more efficient in comparison to the RR designs of Warner [25] and Mangat and Singh [49]. Specially, for this RR design, an unbiased estimator of θ and its variance are given by $\widehat{\theta}_M = \frac{n_1/n-(1-p)}{p}$ with $\mathrm{Var}(\widehat{\theta}_M) = \frac{\lambda_M(1-\lambda_M)}{np^2}$, respectively, where, $\lambda_M = \theta + (1-p)(1-\theta)$. According to the unrelated-question model of Horvitz et al. [39] and the model of Mangat and Singh [49], Chang and Liang [51] conducted a new two-stage unrelated RR design. Gjestvang and Singh [52] adjusted the parameters of the randomization device to propose a more efficient RR model than the models of [25,49,50] to refine the two-stage randomization. Huang [3] used the two-stage RR procedure to improve efficiency of the RR procedure of [25]. Recently, Chang et al. [2] utilized logistic regression to estimate the prevalence of a sensitive feature with a categorical or quantitative explanatory variable.

A new two-stage unrelated RR model was proposed by Vishwakarma et al. [53] to estimate the mean number of individuals in a given population who have a rare sensitive attribute by using Poisson probability distribution, when the proportion of rare non-sensitive unrelated attribute is known and unknown.

3.3. The Generalized Randomized Response Design of Christofides and Some Direct Extensions

In 2003, Christofides [54] provided the generalized RR (GRR) design of a single sensitive question to let respondents have more than two response options and be more protective toward their privacy. It is shown that the GRR design is more efficient in comparison to the RR design of Warner [25]. Let a respondent have one of the sensitive and non-sensitive attributes. If the respondent had the sensitive attribute, let her/him remember the number $L+1$; otherwise, let her/him remember the number 0. Next, she/he utilizes a randomization device to generate a random integer from 1 to L with probability distribution $\boldsymbol{P} = (P_1, P_2, \ldots, P_L)$, where $\sum_{j=1}^{L} P_j = 1$. This number is not reported directly to the surveyor. If the respondent had the sensitive attribute, she/he only provides the answer how far this number is away from $L+1$; otherwise, provide the answer how far this number is away from 0.

Assume that Y_i, $i = 1, 2, \ldots, n$, is respondent i taking the value $L+1$ if having the sensitive attribute and 0 otherwise. T_i is a random integer generated by respondent i using the randomization device to obtain the value j with probability $P_j = P(T_i = j)$, $j = 1, 2, \ldots, L$. Assume that θ is the population proportion of the sensitive trait. Y_i has the Bernoulli distribution with probability $\theta = P(Y_i = L+1)$ and probability $1-\theta = P(Y_i = 0)$, denoted by $Y_i \sim (L+1) \times B(1, \theta)$, where $B(1, \theta)$ denotes the Bernoulli distribution of a random variable taking the value 1 with probability θ and 0 with probability $1-\theta$. See Figure 1 of Lee et al. [55] for illustration of the probability mass functions (pmfs) for Y_i and T_i, respectively. From the GRR design of Christofides [54], the ith respondent reports how far Y_i is away from T_i. Thus, this respondent only provides the value of $D_i = |Y_i - T_i|$, whose pmf is $P(D_i = d) = (1-\theta)P_d + \theta P_{L+1-d}$, $i = 1, 2, \ldots, n$, $d = 1, 2, \ldots, L$. Christofides [54] obtained the expectation of D_i, $\mathrm{E}(D_i) = \mathrm{E}(T_i) + \theta(L+1-2\mathrm{E}(T_i))$, and took $\overline{D} = \sum_{j=1}^{n} D_j/n$ to replace $\mathrm{E}(D_i)$. Because the expectation of T_i is known, $\widehat{\theta}_C = \frac{\overline{D}-\mathrm{E}(T_i)}{L+1-2\mathrm{E}(T_i)}$ is used as an estimator of θ. Similarly, it is easy to verify the variance of D_i as $\mathrm{Var}(D_i) = \mathrm{Var}(T_i) + \theta(1-\theta)(L+1-2\mathrm{E}(T_i))^2$. Hence, Christofides [54] showed $\mathrm{Var}(\widehat{\theta}_C) = \frac{\theta(1-\theta)}{n} + \frac{\mathrm{Var}(T_i)}{n(L+1-2\mathrm{E}(T_i))^2}$. See Appendix A.2. θ can be replaced by $\widehat{\theta}_C$ to obtain an estimator of $\mathrm{Var}(\widehat{\theta}_C)$. The first and second terms of $\mathrm{Var}(\widehat{\theta}_C)$ are the variance because of random sampling and the variance due to the randomization procedure, respectively. If choosing suitable values for $P_1, P_2, \ldots, P_L$ such that $\frac{\mathrm{Var}(T_i)}{n(L+1-2\mathrm{E}(T_i))^2} < \frac{p(1-p)}{n(2p-1)^2}$, then $\widehat{\theta}_C$ is more efficient than $\widehat{\theta}_W$. When $n \to \infty$, $\widehat{\theta}_C$ is asymptotically normally distributed, and, hence, interval estimation can be performed. When $L = 2$, $P_1 = p$ (and $P_2 = 1-p$) and,

hence, this GRR model is reduced to the RR model of Warner [25]. Furthermore, when $L \geq 3$, the mean squared error of $\widehat{\theta}_C$ is smaller in comparison to that of $\widehat{\theta}_W$ [54].

Christofides [54] also showed that $\mathrm{Var}(\widehat{\theta}_C)$ can be reduced by multiple use of the randomization device. In this instance, individual i is asked to use the randomization device m_i times. The m_i repetitions of the procedure must be independent of each other. Let T_{ij} be the number produced by individual i using the randomization device at the jth time. Suppose that $D_{ij} = |Y_i - T_{ij}|$ is the reported number. Define $\widehat{\theta}_{m.} = \frac{\overline{D}_{m.} - \mathrm{E}(T)}{L+1-2\mathrm{E}(T)}$, where $\overline{D}_{m.} = \left(\sum_{i=1}^{n} m_i\right)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m_i} D_{ij}$. Assume that T has the same distribution as the T_{ij}'s, $j = 1, 2, \ldots, m_i$, $i = 1, 2, \ldots, n$. $\widehat{\theta}_{m.}$ is shown to be an unbiased estimator of θ with

$$\mathrm{Var}(\widehat{\theta}_{m.}) = \frac{\sum_{i=1}^{n} {m_i}^2}{\left(\sum_{i=1}^{n} m_i\right)^2}\theta(1-\theta) + \frac{1}{\sum_{i=1}^{n} m_i} \frac{\mathrm{Var}(T)}{[L+1-2\mathrm{E}(T)]^2}.$$

In the special case where when $m_i = m$, $i = 1, 2, \ldots, n$, i.e., each respondent is asked to use the randomization device m times,

$$\mathrm{Var}(\widehat{\theta}_{m.}) = \frac{\theta(1-\theta)}{n} + \frac{\mathrm{Var}(T)}{mn[L+1-2\mathrm{E}(T)]^2}.$$

Thus, when $m_i = m$, $i = 1, 2, \ldots, n$, $\mathrm{Var}(\widehat{\theta}_{m.})$ is then smaller via multiple use of the randomization device in comparison to $\mathrm{Var}(\widehat{\theta}_C) = \frac{\theta(1-\theta)}{n} + \frac{\mathrm{Var}(T)}{n\left[L+1-2\mathrm{E}(T)\right]^2}$ in Christofides [54].

Christofides [54] proposed an improved modification of the RR design of Warner [25] to estimate an unknown proportion of population bearing a sensitive characteristic in a given community. Chaudhuri [56] presented methods to estimate an unknown population proportion of a sensitive attribute when RR data of Christofides [57] are available from unequal probability samples. Christofides [58] extended the GRR model of Christofides [54] to the case of stratified sampling. Christofides [57] extended the GRR model of [54] by proposing an RRTthat allows for estimation of the population proportion of subjects with two sensitive attributes simultaneously. Lee et al. [59] proposed a special model of the GRR version of Christofides [57], called a simple model. They also proposed a so-called crossed model that is more efficient compared to the simple model. Perri et al. [60] applied the crossed model to investigate the phenomena of the induced abortion and illegal immigration simultaneously in Calabria, Italy and also attested to the fact that the crossed model is more efficient.

3.4. Sensitive Characteristics with More Than One Category

It is in the RR model of Warner [25] supposed that every individual in a population is in either the sensitive group or the non-sensitive group, and the population proportion of subjects in the sensitive group is estimated by a survey. Abul-Ela et al. [61] improved the RR design of Warner [25] for the trichotomous population with at least one sensitive group. Hsieh et al. [9] extended the GRR design [54] to the case where there are more than two categories and estimated the proportion of each category by employing the ML method. Hsieh and Lukusa [10] used the ML method and Bayesian approach to estimate the proportion of each group in a trichotomous population. The population with ℓ $(\ell \geq 3)$ related mutually exclusive groups, with at least one and at most $\ell - 1$ of them being sensitive, was also extended by, e.g., Hsieh et al. [9] and Liu and Chow [62]. Recently, Hsieh et al. [12] provided the two-stage multilevel RRT based on an extension of the GRR design in [9] to collect the monthly income data.

3.5. Simultaneous Study of Multiple Sensitive Characteristics

Some works have estimated the population proportion of two sensitive features simultaneously. Barksdale [63] proposed some RRTs to collect data for analysis to investigate two sensitive dichotomous traits. Drane [64] explored the problem of testing independence

between two sensitive dichotomous characteristics by utilizing repeated applications of various RRTs for single attribute. Fox and Tracy [65] estimated the correlation between two sensitive traits. Christofides [57] introduced an RRT to estimate the proportion of subjects with two sensitive attributes simultaneously. Lee et al. [59] extended the RR design in [25] to capture two sensitive characteristics. Afterwards, Ewemooje [66] improved the procedure to estimate the population proportion of two sensitive features at a time by utilizing equal probabilities of protection on the randomization devices. It has been shown that the proposed model is more efficient compared to the model of Lee et al. [59] in some cases. Ewemooje and Amahia [67,68] extended the work of Mangat [50] to propose new and more efficient estimators of the population proportion of respondents bearing two related sensitive traits in survey sampling. Batool and Shabbir [69] considered the problem of estimating the several proportions of two inter-dependent sensitive attributes prevailing in a given population. Xu et al. [70] proposed a new, unique unrelated-question RR model, where each card contains two questions, either both questions on the sensitive characteristics or both questions on the unrelated characteristics. Chung et al. [71] implemented the RRT with multiple sensitive traits and utilized a Bayesian approach to estimate covariance matrices with incomplete information. Chu et al. [72] proposed a new statistical method to combine the RRT, probit modeling, and Bayesian analysis to analyze large-scale online surveys of multiple binary RRs. Recently, Hsieh and Perri [20] provided a logistic regression extension for the RR *simple* and *crossed models* to discuss two related sensitive attributes in [59].

3.6. Randomized Response Techniques for Quantitative Sensitive Data

Greenberg et al. [73] extended the RRT of reducing respondent bias in obtaining answers to sensitive questions from a situation where the response is categorical to that in which the response is quantitative. Gupta et al. [74] estimated the expected mean of the stigmatized variable by using an optional RR sampling. By using double sampling, Grewal et al. [75] estimated the expected mean of a sensitive quantitative variable. Hussain and Shabbir [76] provided an unbiased estimator of the population mean of a sensitive quantitative variable based on multiple selections of numbers from a scrambling distribution to confound the actual response on a sensitive variable with some unrelated variable. Hsieh et al. [12] estimated the personal monthly mean income by using a two-stage multilevel RRT with proportional odds (PO) models.

Hussain et al. [77] proposed a new RR model to estimate the population total of a sensitive variable of quantitative nature. To achieve the objective, they introduced additive scrambling mechanism when sample is drawn through probability proportional to size sampling scheme. Gupta et al. [78] proposed an optional enhanced trust (OET) quantitative RRT model to mitigates the effect of respondents' lack of trust by allowing them who do not trust the traditional additive RRT model to use an alternative scrambling technique. They utilized a combined measure of respondent privacy and model efficiency to demonstrate both theoretically and empirically that the proposed OET model is superior to the traditional model of Warner [79].

3.7. Applications of Randomized Response Techniques to Real Data

Applications of RRTs to real data related to sensitive topics can be found in various works, such as illegitimacy of offspring [40], drug use [72,80–82], incidence of induced abortions [60,62,83,84], fraudulent acts [19,85–88], racism [89,90], sexual behavior [2,9,10,20,55,91–93], cheating in examinations [94], monthly income [12], illegal immigration [95], and conservation [23,24]. In recent years, many researchers have been attracted to using RRTs to collect data on fraudulent behaviors during the COVID-19 pandemic. For example, Mieth et al. [96] used indirect questions to provide prevalence estimates for personal hygiene behavior during the early stages of the COVID-19 pandemic in Germany in 2020. Reiber et al. [97] conducted a survey on intimate partner violence during the COVID-19 pandemic, along with various other studies.

Striegel et al. [98] estimated the prevalence of doping and illicit drug abuse. They used a two-sided z-test to compare the anonymous standardized questionnaire and RRT results with the respective official German National Anti-Doping Agency data on the prevalence of doping. Christiansen et al. [99] measured the prevalence of doping in recreational sport by using the RRT. Mielecka-Kubień and Toniszewski [100] estimated the prevalence of illicit drug use among high school students living in the Silesian voivodship (Poland) by using either the RRTs of forced response design or the Liu-Chow method [101]. Burgstaller et al. [102] argued that the RRT and list experiments would validate and improve prevalence estimates of undeclared work that is defined as a taxable and essentially legal economic activity, but that is not intentionally reported to the relevant authority. They considered an undeclared work case in Germany to demonstrate the strengths and weaknesses of conventional surveys. Furthermore, readers can refer to [22,32] for more studies using IQT for real data.

3.8. Statistical Methods for Randomized Response Data

The two well-known estimation methods, frequentist and Bayesian, in statistics have been applied by several authors RR data.

3.8.1. Frequentist Methods

After collecting data through RR designs, estimation and statistical inference of unknown population parameters of interest, such as the proportion of sensitive characteristics, honest response rate, and sensitivity level of questions, can be carried out. In the frequentist approach, the commonly used classical methods are the ML method and method of moments (MM). A common problem with these two methods is that the estimated parameter value may be out of the true parameter space. For example, the estimate of the proportion of a sensitive feature may fall outside the interval $[0, 1]$; see, e.g., [20,33,103]. In addition, the calculation of ML estimates is sometimes more complicated and requires numerical methods. The expectation–maximization (EM) method [104] can be used to address this issue; see, e.g., [22,105–107]. Specifically, Bourke and Moran [105] presented the particular applicability of the EM algorithm in obtaining ML estimates of proportions where the sensitive data are collected by using an RR design. They considered two kinds of RR designs: related-question [25] and unrelated-question [40] designs. van den Hout and Kooiman [107] developed a fast and straightforward EM algorithm to obtain ML estimates of the parameters of a linear regression model with categorical covariates subject to RR. Groenitz [22] derived a general EM algorithm to obtain general ML estimates of the parameters of a logistic regression model. Recently, to obtain an efficient estimator of the proportion of a sensitive characteristic and to investigate the association between the sensitive characteristic or latent variable and an observed binary variable, Lee et al. [106] proposed a combination of Warner's RRT [25] and a latent class model. An EM algorithm is proposed to estimate the model parameters. However, the EM method also has its own weaknesses, such as its tendency to fail to converge to the true value; see, e.g., [108].

3.8.2. Bayesian Method

Some authors have suggested using the Bayesian method to deal with the weaknesses and improve the efficiency of previous estimation methods in cases where some prior information on parameters is available. The major references on the RRT in the Bayesian framework are listed below. Winkler and Franklin [84] proposed a seminal work in which the Bayesian approach was used to analyze RR data. Hussain et al. [109], Migon and Tachibana [110], and a bunch of other authors then used the Bayesian method to estimate the population proportion of a sensitive trait in Warner's RR design [25]. Pitz [111] used a Bayesian analysis of the model of Fidler and Kleiknecht [112] to give a more useful estimation when the sample size is not large or the response proportions are extreme. O'hagan [113] employed a non-parametric approach to derive Bayes linear estimators.

Oh [114] and Unnikrishnan and Kunte [115] used the Bayesian method through a Gibbs sampling algorithm to estimate parameters of interest by introducing latent variables to an RR model. Bar-lev et al. [103] presented a common conjugate prior structure for some RR models. Hussain and Shabbir [116] used a stratified random sampling protocol and the Bayesian method to estimate the population proportion of a sensitive feature. Song and Kim [117] addressed the Bayesian formulation of two types of Poisson regression models for RR sum score variables under the self-protection assumption. Adepetun and Adewara [118] utilized both Kumaraswamy and generalised beta prior distributions to propose the Bayesian estimators of the population proportion of a stigmatized characteristic when data were obtained via the RRT of Kim and Warde [43]. Groenitz [119] proposed a design method for multiple-choice sensitive features and provided the Bayesian method combined with Gibbs sampling and Markov chain Monte Carlo (MCMC) to estimate the population proportions of multichotomous sensitive features. Song and Kim [120] employed the RRT to propose a Bayesian estimation of the rate of a rare sensitive trait. Mehta and Aggarwal [4] and Narjis and Shabbir [5] provided Bayesian estimation of a sensitivity level and the population proportion of a sensitive attribute of optional unrelated-question RR models.

Recently, Nandram and Yu [108] introduced a Bayesian analysis of spare counts gathered from the unrelated-question design. More recently, Hsieh and Lukusa [10] implemented a Bayesian framework for multilevel RR data and compared the Bayesian method with the ML method for estimating the population proportion of individuals aged 18–54 years who self-reported as bisexual and homosexual among Taiwanese. Hsieh and Perri [20] proposed a Gibbs sampling algorithm to estimate the population proportion of the sensitive characteristic θ. They compared, in connection with the GRR data-collection model of [54], the MM, ML, and Bayesian methods for the estimation of the population proportion of non-heterosexuals aged 20 years or older for the Taiwanese population, gender groups, and age groups. Specifically, suppose that $\{(D_i, Y_i) : i = 1, 2, \ldots, n\}$ are available. The joint pmf of (D_i, Y_i) is given by $P(D_i = d_i, Y_i = y_i) = \left(P_{L+1-d_i}\theta\right)^{I(y_i=L+1)}\left(P_{d_i}(1-\theta)\right)^{I(y_i=0)}$, where $I(\cdot)$ is an indicator function. Given $\mathcal{D}^* = \{(d_i, y_i) : i = 1, 2, \ldots, n\}$, the likelihood function can be obtained as $\mathcal{L}^*(\theta|\mathcal{D}^*) = \theta^{\sum_{i=1}^n I(y_i=L+1)}(1-\theta)^{\sum_{i=1}^n I(y_i=0)}\prod_{i=1}^n P_{L+1-d_i}^{I(y_i=L+1)}P_{d_i}^{I(y_i=0)}$. Thus, given that a beta prior distribution with parameters α_1 and α_2, denoted by $\theta \sim \text{Beta}(\alpha_1, \alpha_2)$, is assigned to θ, [20] derived the conditional posterior distribution of θ given $\mathcal{D}^*$ as $\theta|\mathcal{D}^* \sim \text{Beta}(\alpha_1 + \sum_{i=1}^n I(y_i = L+1), \alpha_2 + \sum_{i=1}^n I(y_i = 0))$. However, in practice, through the GRR design of [54], only $\boldsymbol{d} = (d_1, d_2, \ldots, d_n)$ can be obtained, so, [20] treated $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ as latent variables to derive the conditional distribution of Y_i given θ and $D_i = d_i$. The probability of $Y_i = L+1$ given θ and $D_i = d_i$ is $p(\theta, d_i) = P\left(Y_i = L+1|\theta, D_i = d_i\right) = \frac{P_{L+1-d_i}\theta}{P_{L+1-d_i}\theta + P_{d_i}(1-\theta)}$, $i = 1, 2, \ldots, n$. The conditional distribution of Y_i given θ and $D_i = d_i$ is then a Bernoulli distribution with probability $p(\theta, d_i)$ of $Y_i = L+1$ and probability $1 - p(\theta, d_i)$ of $Y_i = 0$, denoted by $Y_i|\theta, D_i = d_i \sim (L+1) \times B(1, p(\theta, d_i))$, $i = 1, 2, \ldots, n$.

Chung et al. [71] used a Bayesian approach to estimate covariance matrices with incomplete information in a population with multiple sensitive characteristics. According to the idea of Hsieh and Perri [20], Lee et al. [55] used the Bayesian estimation method through data augmentation and MCMC to estimate the prevalence of the population possessing the sensitive attribute and the distribution of a categorical or quantitative variable in each of the non-sensitive and sensitive groups. The deviance information criterion and marginal likelihood are employed to select a suitable model to describe the association of the sensitive characteristic with the auxiliary random variable in this work. Chu et al. [72] combined the RRT, probit modeling, and Bayesian approach to analyze large-scale online surveys of multiple binary RRs.

In 2023, Ewemooje et al. [82] proposed a new Bayesian estimation method for Alternative Tripartite RRTs to gain the proportion of individuals belonging to a sensitive character. The proposed Bayesian estimators used the Kumaraswamy and the generalized beta prior

distributions. A comparison of the classical technique and Bayesian method is provided in [82].

3.9. Use of Auxiliary Information in Randomized Response Problems

3.9.1. Regression Models for Randomized Response Data

In sample surveys on sensitive topics, besides sensitive information of interest collected by IQTs, information on some auxiliary variables is also obtained. The data of these auxiliary variables are collected by using direct questioning techniques (DQTs). Using these auxiliary variables reasonably to improve computational efficiency is an important issue that has received the attention of several authors. The following is a brief summary of the use of auxiliary variables in sensitive variable research.

In 1983, Maddala [81] employed a logit model to investigate the relationships between auxiliary variables and randomized response survey data through the RR design of Warner [25]. The author obtained ML model parameter estimates using the Newton–Raphson iterative procedure. An estimate of the asymptotic covariance matrix was shown. This logit model was then illustrated for the first time in real data by Kerkvliet [121] in the study of college students' cocaine use at two public universities in the United States that were surveyed in 1989. Scheers and Dayton [94] established a theory for an extension of the RR design of [25] and a covariate extension of the unrelated-question RR design of Greenberg et al. [40]. They showed that if the relationship between the covariates and the sensitive population proportions is correctly specified, the covariate RR model is relatively more efficient. In 1996, van der Heijden and van Gils [87] presented the model where the response variable is subject to the RR design of Boruch [122] or Kuk [123]. van den Hout et al. [88] discussed univariate and multivariate logistic regression where response variables are subject to RR.

van den Hout and Kooiman [107] derived the likelihood of the linear regression model with categorical covariates subject to RR. They developed a fast and straightforward EM algorithm to obtain ML estimates of the regression parameters. Cruyff et al. [124] provided a review of regression procedures for RR data, including the univariate and multivariate logistic regression models, PO regression model, item response model, and self-protective responses. Blair et al. [29] presented how their developed multivariate logistic regression techniques were employed to analyze data collected from the four RR designs: mirrored question, forced response, disguised response, and unrelated question. Hsieh et al. [85,86] and recently Chang et al. [2] estimated the prevalence of a sensitive characteristic with a categorical or quantitative explanatory variable by fitting logistic regression.

Let Y be the answer to a sensitive question, $\boldsymbol{Z}$ a vector of covariates that are always observed, and $\boldsymbol{X}$ another covariate vector that may be missing on some subjects. Assume that $\boldsymbol{W}$ is a surrogate for $\boldsymbol{X}$ and independent of Y given $\boldsymbol{X}$ and $\boldsymbol{Z}$. Let $Y = 1$ and $Y = 0$ denote answering "Yes" and "No", respectively, to the sensitive question. Now consider the following logistic regression model:

$$P(Y = 1|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) = H(\beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{X} + \boldsymbol{\beta}_2^T \boldsymbol{Z}) = H(\boldsymbol{\beta}^T \mathcal{X}),$$

where $H(u) = 1/(1 + \exp(-u))$ and $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$ is a vector of unknown parameters for $\mathcal{X} = (1, \boldsymbol{X}^T, \boldsymbol{Z}^T)^T$. Under the RRT, Y is not observable. Let Y^0 denote the binary response to the sensitive question based on some RRT, such as [25,40,42,49]. The probability of Y^0 given $\boldsymbol{X}$ and $\boldsymbol{Z}$ can then be expressed as follows:

$$P(Y^0 = 1|\boldsymbol{X}, \boldsymbol{Z}) = kH(\boldsymbol{\beta}^T \mathcal{X}) + s, \tag{1}$$

where k and s are known constants in different RRTs. For example, $k = 2p - 1$ and $s = 1 - p$ in Warner's RRT [25]; $k = p$ and $s = (1 - p)c$ in the RRT proposed by Greenberg et al. [40], where p is the probability of selecting the sensitive question and c is the probability of selecting the innocuous question to answer "Yes".

Most recently, Groenitz [22] used logistic regression for the analysis of direct data on the covariates and indirect data on the sensitive variable. The author derived a general algorithm for the ML estimation and a general procedure for variance estimation. Ronning [125] analyzed effects of RR with respect to some binary dependent variable on the estimation of the probit model. Hsieh and Perri [95] proposed a logistic regression extension for analyzing the factors that influence two sensitive variables when data are collected by the RR simple and crossed models.

3.9.2. Missing Data in Randomized Response Problems

Most works on RR data assume that the data are observable. That means the data used in these works are assumed to be fully observed. This assumption is sometimes difficult to achieve in practice. Hsieh et al. [85] developed two semiparametric approaches to estimate the parameters of logistic regression for RR data with missing covariates. After that, Hsieh et al. [86] utilized a logistic regression model for analyzing RR data with covariates missing at random (MAR). Hsieh et al. [13] combined the unrelated-question RRT of Greenberg et al. [40] and the related-question RRT of Warner [25] to address the issue of an innocuous question in the unrelated-question RR design. They utilized logistic regression with missing data to estimate the prevalence of the sensitive characteristic. Lee et al. [14] combined both the unrelated-question RRT of [40] and the DQT under a missing data setting to propose a data-collection method for surveys of sensitive issues. Recently, Hsieh et al. [12] employed PO regression on the two-stage multilevel RRT of [9] to investigate the monthly income when some covariates are MAR.

Let δ indicate whether $\boldsymbol{X}$ is observed ($\delta = 1$) or not ($\delta = 0$). Assume that $\boldsymbol{W}$ is a possible surrogate of $\boldsymbol{X}$ such that $\boldsymbol{W}$ is dependent on $\boldsymbol{X}$ and independent of Y^0 given $\boldsymbol{X}$ and $\boldsymbol{Z}$. Hsieh et al. [85,86] assumed that the missing mechanism is missing at random (MAR) [126], i.e., the probability of $\boldsymbol{X}$ being observed, the selection probability $P(\delta = 1|Y^0, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) = \pi(Y^0, \boldsymbol{Z}, \boldsymbol{W})$, depends on $(Y^0, \boldsymbol{Z}, \boldsymbol{W})$, but not on $\boldsymbol{X}$. The validation data set consists of $\{(Y_i^0, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{W}_i, \delta_i = 1) : i = 1, 2, \ldots, n\}$, and the non-validation data set includes $\{(Y_i^0, \boldsymbol{Z}_i, \boldsymbol{W}_i, \delta_i = 0) : i = 1, 2, \ldots, n\}$. Let $v_1, v_2, \ldots, v_g$ denote the distinct values of the $\boldsymbol{V}_i$'s, where $\boldsymbol{V}_i = (\boldsymbol{Z}_i, \boldsymbol{W}_i)$. For $\boldsymbol{v} \in \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_g\}$ and $y^0 = 0, 1$, $\pi(y^0, \boldsymbol{v})$ is estimated by

$$\widehat{\pi}(y^0, \boldsymbol{v}) = \frac{\sum_{i=1}^{n} \delta_i I(Y_i^0 = y^0, \boldsymbol{V}_i = \boldsymbol{v})}{\sum_{i=1}^{n} I(Y_i^0 = y^0, \boldsymbol{V}_i = \boldsymbol{v})}.$$

To estimate $\boldsymbol{\beta}$, Hsieh et al. [86] proposed the Horvitz and Thompson-type weighted estimating equations [127] as follows:

$$\boldsymbol{U}_w(\boldsymbol{\beta}, \widehat{\boldsymbol{\pi}}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{\delta_i}{\widehat{\pi}(Y_i^0, \boldsymbol{V}_i)} \mathcal{X}_i A_i(\boldsymbol{\beta}) \left[Y_i^0 - [kH(\boldsymbol{\beta}^T \mathcal{X}_i) + s] \right] \right\} = \boldsymbol{0},$$

where $\widehat{\boldsymbol{\pi}} = (\widehat{\pi}(Y_1^0, \boldsymbol{V}_1), \widehat{\pi}(Y_2^0, \boldsymbol{V}_2), \ldots, \widehat{\pi}(Y_n^0, \boldsymbol{V}_n))$,

$$A_i(\boldsymbol{\beta}) = \frac{kH(\boldsymbol{\beta}^T \mathcal{X}_i)[1 - H(\boldsymbol{\beta}^T \mathcal{X}_i)]}{(kH(\boldsymbol{\beta}^T \mathcal{X}_i) + s)(1 - kH(\boldsymbol{\beta}^T \mathcal{X}_i) - s)}. \tag{2}$$

Hsieh et al. [86] also proposed to model $\pi(Y_i^0, \boldsymbol{V}_i)$ with logistic regression with known parameters or unknown parameters to discuss the efficiency problem.

Multiple imputation (MI) is another statistical technique to deal with the missing data. Lee et al. [128] and Stoklosa et al. [129] proposed generating imputed data by applying the MI scheme developed by Wang and Chen [130] in different areas. One can estimate

the parameters of the RR regression model in (1) by utilizing the empirical conditional distribution function (CDF) as follows:

$$\widehat{F}(\boldsymbol{x}|Y_i^0, \boldsymbol{V}_i) = \sum_{r=1}^{n} \left\{ \frac{\delta_r I(Y_r^0 = Y_i^0, \boldsymbol{V}_r = \boldsymbol{V}_i)}{\sum_{j=1}^{n} I(Y_j^0 = Y_i^0, \boldsymbol{V}_j = \boldsymbol{V}_i)} \right\} I(\boldsymbol{X}_r \leq \boldsymbol{x}). \tag{3}$$

A unified estimate for the MI procedure proposed by Rubin [131] is the average of estimates obtained from all imputed data sets. Given the number of imputations M, the MI approach is summarized as follows:

Step 1. For missing $\boldsymbol{X}_i$ $(\delta_i = 0)$, generate $\widetilde{\boldsymbol{X}}_{vi}$ from the empirical CDF $\widehat{F}(\boldsymbol{x}|Y_i^0, \boldsymbol{V}_i)$ in (3), $v = 1, 2, \ldots, M$.

Step 2. Let $\widehat{\boldsymbol{\beta}}_v$ denote the solution to the following estimating equations:

$$\begin{aligned} \boldsymbol{U}_v(\boldsymbol{\beta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big\{ & \delta_i \mathcal{X}_i A_i(\boldsymbol{\beta}) \Big[Y_i^0 - [kH(\boldsymbol{\beta}^T \mathcal{X}_i) + s] \Big] \\ & + (1 - \delta_i) \widetilde{\mathcal{X}}_{vi} \widetilde{A}_{vi}(\boldsymbol{\beta}) \Big[Y_i^0 - [kH(\boldsymbol{\beta}^T \widetilde{\mathcal{X}}_{vi}) + s] \Big] \Big\} \\ = \mathbf{0}, & \end{aligned} \tag{4}$$

where $\widetilde{\mathcal{X}}_{vi} = (1, \boldsymbol{X}_{vi}^T, \boldsymbol{Z}_i^T)^T$ and is used to replace $\mathcal{X}_i$ in $A_i(\boldsymbol{\beta})$ in (2) to denote $\widetilde{A}_{vi}(\boldsymbol{\beta})$.

Step 3. The MI estimate of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}}_{m1} = \sum_{v=1}^{M} \widehat{\boldsymbol{\beta}}_v / M$.

Lee et al. [128] provided the second MI-type method as in Fay [132] to estimate $\boldsymbol{\beta}$. In step 2, one can define the following estimating function:

$$\boldsymbol{U}_{m2}(\boldsymbol{\beta}) = \frac{1}{M} \sum_{v=1}^{M} \boldsymbol{U}_v(\boldsymbol{\beta}).$$

Let $\widehat{\boldsymbol{\beta}}_{m2}$ denote the solution to the estimating equations $\boldsymbol{U}_{m2}(\boldsymbol{\beta}) = \mathbf{0}$. The asymptotic properties of the two MI estimators, $\widehat{\boldsymbol{\beta}}_{m1}$ and $\widehat{\boldsymbol{\beta}}_{m2}$, and their corresponding asymptotic variance estimators still need to be established.

In the above discussion, we considered all the elements of $\boldsymbol{X}$ to be missing simultaneously. In practice, the elements of $\boldsymbol{X}$ may be missing simultaneously or separately in the RR regression model. Now consider $\boldsymbol{X}_i = (\mathtt{X}_{1i}^T, \mathtt{X}_{2i}^T)^T$, where $\mathtt{X}_{1i}$ and $\mathtt{X}_{2i}$ may be missing simultaneously or separately.

Define the missingness statuses of the data as follows. For $i = 1, 2, \ldots, n$, $\delta_{i1} = 1$ if both $\mathtt{X}_{1i}$ and $\mathtt{X}_{2i}$ are observed; 0 otherwise. $\delta_{i2} = 1$ if $\mathtt{X}_{1i}$ is missing and $\mathtt{X}_{2i}$ is observed; 0 otherwise. $\delta_{i3} = 1$ if $\mathtt{X}_{1i}$ is observed and $\mathtt{X}_{2i}$ is missing; 0 otherwise. $\delta_{i4} = 1$ if both $\mathtt{X}_{1i}$ and $\mathtt{X}_{2i}$ are missing; 0 otherwise. Assume that $\mathtt{W}_1$ and $\mathtt{W}_2$ are the possible surrogates of $\mathtt{X}_1$ and $\mathtt{X}_2$, respectively, such that $\mathtt{W}_1$ and $\mathtt{W}_2$ are dependent on $\mathtt{X}_1$ and $\mathtt{X}_2$ and independent of Y^0 given $\boldsymbol{X}$ and $\boldsymbol{Z}$. Let $\boldsymbol{W} = (\mathtt{W}_1^T, \mathtt{W}_2^T)^T$. Under the assumption of MAR mechanism [126] of $\mathtt{X}_1$ and $\mathtt{X}_2$, the selection probability model is assumed as follows:

$$P(\delta_{ij} = 1|Y_i^0, \mathtt{X}_{1i}, \mathtt{X}_{2i}, \boldsymbol{Z}_i, \boldsymbol{W}_i) = \pi_j(Y_i^0, \boldsymbol{V}_i),\ j = 1, 2, 3, 4, \tag{5}$$

where $\boldsymbol{V}_i = (\boldsymbol{Z}_i^T, \boldsymbol{W}_i^T)^T$ and $\sum_{j=1}^{4} \pi_j(Y_i^0, \boldsymbol{V}_i) = 1$. $\pi_j(Y_i^0, \boldsymbol{V}_i)$'s are the nuisance parameters and unknown, although it may be specified at design stage in some applications.

Lee et al. [133] proposed two different types of MI methods for the estimation of the parameters of the logistic regression model with covariates missing separately or simultaneously. Their approaches, which are based on the ideas of [130,132], involve a two-step procedure instead of the three-step procedure as in the traditional MI approaches, in order to reduce the computing time, and are more efficient in estimation. These estimation methods can also be applied to the RRT. For example, one can use the first approach of [133] in the RRT below.

Consider the following empirical CDFs of $\mathtt{X}_{1i}$, given $(\mathtt{X}_{2i}, Y_i^0, \boldsymbol{V}_i)$, $\mathtt{X}_{2i}$, given $(\mathtt{X}_{1i}, Y_i^0, \boldsymbol{V}_i)$, and $\boldsymbol{X}_i$ given $(Y_i^0, \boldsymbol{V}_i)$:

$$\widetilde{F}_{\mathtt{X}_{1i}}(\mathtt{x}_1|\mathtt{X}_{2i}, Y_i^0, \boldsymbol{V}_i) = \sum_{k=1}^{n}\left(\frac{\delta_{k1}I(Y_k^0 = Y_i^0, \mathtt{X}_{2k} = \mathtt{X}_{2i}, \boldsymbol{V}_k = \boldsymbol{V}_i)}{\sum_{s=1}^{n}\delta_{s1}I(Y_s^0 = Y_i^0, \mathtt{X}_{2s} = \mathtt{X}_{2i}, \boldsymbol{V}_s = \boldsymbol{V}_i)}\right)I(\mathtt{X}_{1k} \le \mathtt{x}_1),$$

$$\widetilde{F}_{\mathtt{X}_{2i}}(\mathtt{x}_2|\mathtt{X}_{1i}, Y_i^0, \boldsymbol{V}_i) = \sum_{k=1}^{n}\left(\frac{\delta_{k1}I(Y_k^0 = Y_i^0, \mathtt{X}_{1k} = \mathtt{X}_{1i}, \boldsymbol{V}_k = \boldsymbol{V}_i)}{\sum_{s=1}^{n}\delta_{s1}I(Y_s^0 = Y_i^0, \mathtt{X}_{1s} = \mathtt{X}_{1i}, \boldsymbol{V}_s = \boldsymbol{V}_i)}\right)I(\mathtt{X}_{2k} \le \mathtt{x}_2),$$

$$\widetilde{F}_{\boldsymbol{X}_i}(\boldsymbol{x}|Y_i^0, \boldsymbol{V}_i) = \sum_{k=1}^{n}\left(\frac{\delta_{k1}I(Y_k^0 = Y_i^0, \boldsymbol{V}_k = \boldsymbol{V}_i)}{\sum_{s=1}^{n}\delta_{s1}I(Y_s^0 = Y_i^0, \boldsymbol{V}_s = \boldsymbol{V}_i)}\right)I(\boldsymbol{X}_k \le \boldsymbol{x}),$$

respectively. The two steps of the MI method are given as follows:

Step 1. ***Imputation:*** Generate the vth imputed ("completed") data set, $v = 1, 2, \ldots, M$, based on the missingness status of $\boldsymbol{X}_i = (\mathtt{X}_{1i}^T, \mathtt{X}_{2i}^T)^T$, $i = 1, 2, \ldots, n$.

(i) If $\delta_{i1} = 1$, keep the values of $\mathtt{X}_{1i}$ and $\mathtt{X}_{2i}$, and define $\mathcal{X}_i = (1, \mathtt{X}_{1i}^T, \mathtt{X}_{2i}^T, \boldsymbol{Z}_i^T)^T$ for all v.

(ii) If $\delta_{i2} = 1$, keep the value of $\mathtt{X}_{2i}$, and generate $\widetilde{\mathtt{X}}_{1iv}$ from $\widetilde{F}_{\mathtt{X}_{1i}}(\mathtt{x}_1|\mathtt{X}_{2i}, Y_i^0, \boldsymbol{V}_i)$ to impute the missing value of $\mathtt{X}_{1i}$, and define $\widetilde{\mathcal{X}}_{2iv} = (1, \widetilde{\mathtt{X}}_{1iv}^T, \mathtt{X}_{2i}^T, \boldsymbol{Z}_i^T)^T$.

(iii) If $\delta_{i3} = 1$, keep the value of $\mathtt{X}_{1i}$, and generate $\widetilde{F}_{\mathtt{X}_{2i}}(\mathtt{x}_2|\mathtt{X}_{1i}, Y_i^0, \boldsymbol{V}_i)$ to impute the missing value of $\mathtt{X}_{2i}$, and define $\widetilde{\mathcal{X}}_{3iv} = (1, \mathtt{X}_{1i}^T, \widetilde{\mathtt{X}}_{2iv}^T, \boldsymbol{Z}_i^T)^T$.

(iv) If $\delta_{i4} = 1$, generate $\widetilde{\mathtt{X}}_{1iv}$ and $\widetilde{\mathtt{X}}_{2iv}$ from $\widetilde{F}_{\boldsymbol{X}_i}(\boldsymbol{x}|Y_i^0, \boldsymbol{V}_i)$ to impute the missing values of $\mathtt{X}_{1i}$ and $\mathtt{X}_{2i}$, and define $\widetilde{\mathcal{X}}_{4iv} = (1, \widetilde{\mathtt{X}}_{1iv}^T, \widetilde{\mathtt{X}}_{2iv}^T, \boldsymbol{Z}_i^T)^T$.

Step 2. ***Analysis:*** Solve the following estimating equations:

$$\boldsymbol{U}_M(\boldsymbol{\beta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\delta_{i1}\mathcal{X}_i A_i(\boldsymbol{\beta})\Big(Y_i^0 - [kH(\boldsymbol{\beta}^T\mathcal{X}_i) + s]\Big) + \frac{1}{M}\sum_{j=2}^{4}\sum_{v=1}^{M}\delta_{ij}\widetilde{\mathcal{X}}_{jiv}\widetilde{A}_{jiv}(\boldsymbol{\beta})\Big(Y_i^0 - [kH(\boldsymbol{\beta}^T\widetilde{\mathcal{X}}_{jiv}) + s]\Big)\Big\} = \boldsymbol{0},$$

to obtain the MI estimate of $\boldsymbol{\beta}$, where

$$A_i(\boldsymbol{\beta}) = \frac{kH^{(1)}(\boldsymbol{\beta}^T\mathcal{X}_i)}{[kH(\boldsymbol{\beta}^T\mathcal{X}_i) + s][1 - kH(\boldsymbol{\beta}^T\mathcal{X}_i) - s]},$$

$$\widetilde{A}_{jiv}(\boldsymbol{\beta}) = \frac{kH^{(1)}(\boldsymbol{\beta}^T\widetilde{\mathcal{X}}_{jiv})}{[kH(\boldsymbol{\beta}^T\widetilde{\mathcal{X}}_{jiv}) + s][1 - kH(\boldsymbol{\beta}^T\widetilde{\mathcal{X}}_{jiv}) - s]},$$

with $H^{(1)}(\cdot) = H(\cdot)[1 - H(\cdot)]$. In Step 1, the aforementioned empirical CDFs are utilized to generate imputed data sets by using the complete-case data. δ_{ij}s are employed to identify exactly the partitioned covariate vector without missing observations that are used as the information for the empirical CDFs. More specifically, when $\delta_{i2} = 1$ ($\delta_{i3} = 1$), one can employ the condition from the observed $\mathtt{X}_{2i}$ ($\mathtt{X}_{1i}$), Y_i^0, and $\boldsymbol{V}_i$ to generate a set of values to impute the missing values of $\mathtt{X}_{1i}$ ($\mathtt{X}_{2i}$). When $\delta_{i4} = 1$, i.e., $\mathtt{X}_{1i}$ and $\mathtt{X}_{2i}$ missing simultaneously, the condition from Y_i^0 and $\boldsymbol{V}_i$ is utilized to generate a set of values to impute the missing values of $\mathtt{X}_{1i}$ and $\mathtt{X}_{2i}$. Therefore, the estimation is more efficient. The estimation method can reduce computing time because it only uses two steps to solve the estimating equations once. The asymptotic properties of the MI estimators need to be established, along with the estimation of their variances.

3.9.3. Investigation of Influence of a Sensitive Trait on a Non-Sensitive Variable

In general, the aforementioned studies evaluate the influences of auxiliary variables on sensitive variables of interest. However, there has not been any work evaluating the association of a sensitive variable with auxiliary variables of interest, i.e., whether or not

some random variable of interest on the research subjects depends on the sensitive characteristic. Therefore, motivated by the issue, Lee et al. [55] proposed mixture models for assessing the dependency relationship. Auxiliary information includes a univariate categorical variable, a univariate quantitative variable, and a multivariate quantitative variable to examine in turn. They proposed the Bayesian method through data augmentation and MCMC to estimate the prevalence of the population possessing the sensitive feature and the distribution of a categorical or quantitative variable in each of the non-sensitive and sensitive groups. Moreover, they employed three Bayesian model selection criteria to choose the most suitable one among the proposed models to explore the association of the sensitive variable with a multivariate auxiliary variable in simulation studies. Finally, the two Bayesian model selection criteria, deviance information criterion [134], and marginal likelihood [135] were utilized to choose a more suitable model for the univariate auxiliary variable case.

It is difficult to study empirically sexual behaviors due to their sensitive nature. Accurate estimation of the prevalence and frequency of sexual behaviors is difficult using standard techniques, refer to, e.g., [20,92]. There are various works analyzing the efficacy of the RRT, and more generally IQTs, to accomplish honest self-reporting about sexual behaviors, compared to traditional survey techniques. Refer to, e.g., [92], for more detailed discussions. The sensitive issue of a one-night stand was also mentioned in some materials, including, e.g., Wentland and Reissing [136] and Kaspar et al. [137]. However, the number of research works on this behavior is quite modest. A study on this topic can help researchers, managers, and society have a more complete view of this sexual behavior of young people. Lee et al. [55] applied their proposed methodology to study the influence of the response to the sensitive question, "Have you ever had a one-night stand through a dating site or mobile app?", on each of the response to the statement, "I am considering finding a one-night stand through a dating site or mobile app", the response to the question, "How many significant others have you had?", and "the sum of scores of responses to six internet dating experience questions" by using the data set collected from the survey study of sexuality of freshmen at Feng Chia University in Taiwan in 2016.

Recently, Lee et al. [106] proposed a combination of Warner's RRT [25] and a latent class model to provide a more efficient estimation of the proportion of a sensitive characteristic and to investigate the association between the sensitive characteristic or latent variable and an observed binary variable. The concept of the relationship between the sensitive characteristic variable and other variables in [55] was extended by employing the RR design of [25] to collect sensitive characteristic information. Let Y be the answer to the sensitive question and $\boldsymbol{Z}$ an observed vector of k dichotomous variables with values 0 and 1. In Warner's RRT [25], p is the probability of selecting the sensitive question and Y^0 is a binary outcome, where $Y^0 = 1$ and $Y^0 = 0$ denote answering "Yes" and "No", respectively. Based on a latent variable model, it is assumed that under the group to which an individual is known to belong, the corresponding observed/manifest variables are independent. Therefore, assume that $Z_1, Z_2, \ldots, Z_k$ given Y are independent. Let $P(Y = 1) = \theta$ and $P(Z_s = 1|Y = y) = \alpha_{ys}$, $y = 0, 1$, $s = 1, 2, \ldots, k$. Lee et al. [106] provided the joint probability distribution of $Y^0, Z_1, Z_2, \ldots, Z_k$ as follows:

$$\begin{aligned}
&P(Y^0 = 1, Z_1 = z_1, Z_2 = z_2, \ldots, Z_k = z_k) \\
&= P(Y^0 = 1, Z_1 = z_1, Z_2 = z_2, \ldots, Z_k = z_k|Y = 1)P(Y = 1) \\
&\quad + P(Y^0 = 1, Z_1 = z_1, Z_2 = z_2, \ldots, Z_k = z_k|Y = 0)P(Y = 0) \\
&= p\theta \prod_{s=1}^{k} \alpha_{1s}^{z_s}(1 - \alpha_{1s})^{1-z_s} + (1 - p)(1 - \theta) \prod_{s=1}^{k} \alpha_{0s}^{z_s}(1 - \alpha_{0s})^{1-z_s}
\end{aligned}$$

and

$$\begin{aligned}
&P(Y^0 = 0, Z_1 = z_1, Z_2 = z_2, \ldots, Z_k = z_k) \\
&= P(Y^0 = 0, Z_1 = z_1, Z_2 = z_2, \ldots, Z_k = z_k | Y = 1) P(Y = 1) \\
&\quad + P(Y^0 = 0, Z_1 = z_1, Z_2 = z_2, \ldots, Z_k = z_k | Y = 0) P(Y = 0) \\
&= (1-p)\theta \prod_{s=1}^{k} \alpha_{1s}^{z_s} (1-\alpha_{1s})^{1-z_s} + p(1-\theta) \prod_{s=1}^{k} \alpha_{0s}^{z_s} (1-\alpha_{0s})^{1-z_s}.
\end{aligned}$$

For obtaining the RR data and k-variate dichotomous data of responses to DQ, Ref. [106] proposed an EM algorithm to estimate θ, α_{1s}, and α_{0s}, $s = 1, 2, \ldots, k$. They estimated the variances of estimators using the bootstrap method. An analytic expression for the asymptotic variance still needs to be established. However, the k-variate data of response to DQ are often not dichotomous. For example, "I think online dating is very new/modern" is DQ, and there are five response options: "very consistent", "almost consistent", "fairly consistent", "a bit consistent" and "very inconsistent". Therefore, one can extend the case of k-variate dichotomous responses to DQ in [106] to the case of k-variate multiple responses to DQ. Define $P(Z_s = r | Y = y) = \alpha_{ys,r}$, $r = 1, 2, \ldots, B_s$, with $\sum_{r=1}^{B_s} \alpha_{ys,r} = 1$, where $y = 0, 1$. Under the assumption that $Z_1, Z_2, \ldots, Z_k$ given Y are independent, one can express the joint probability distribution of $Y^0, Z_1, Z_2, \ldots, Z_k$ as follows:

$$P(Y^0 = 1, Z_1 = z_1, Z_2 = z_2, \ldots, Z_k = z_k) = p\theta \prod_{s=1}^{k} \prod_{r=1}^{B_s} \alpha_{1s,r}^{I(z_s = r)} + (1-p)(1-\theta) \prod_{s=1}^{k} \prod_{r=1}^{B_s} \alpha_{0s,r}^{I(z_s = r)}$$

and

$$P(Y^0 = 0, Z_1 = z_1, Z_2 = z_2, \ldots, Z_k = z_k) = (1-p)\theta \prod_{s=1}^{k} \prod_{r=1}^{B_s} \alpha_{1s,r}^{I(z_s = r)} + p(1-\theta) \prod_{s=1}^{k} \prod_{s=1}^{B_s} \alpha_{0s,r}^{I(z_s = r)}.$$

To estimate these parameters $\theta, \alpha_{1s,r}$, and $\alpha_{0s,r}$, $s = 1, 2, \ldots, k$, $r = 1, 2, \ldots, B_s$, a procedure must be developed. One can consider an EM algorithm or the Newton–Raphson method to solve unbiased estimating equations for these parameters and, hence, estimate the variances of their estimators. Another way to estimate these parameters is to use the Bayesian approach, which involves combining the MCMC/Gibbs sampler to generate samples from the posterior distribution of these parameters.

3.10. Statistical Software: Packages and Modules for Randomized Response Data

Some authors have used statistical software to perform analysis of data from randomized surveys. For instance, Hox and Lensvelt-Mulders [138] presented a way to analyze the relations between RR estimates and explanatory variables by using standard structural equation modeling software, `Mplus`. Sehra [139] provided `SAS` code to perform analysis of data gathered from a two-stage additive optional RR model. Jann [140] presented the `Stata` module **rrlogit** to fit logistic regression to RR data. `R` software is also commonly used in RR data analysis. Tian and Tang [31] provided numerous `R` programs to illustrate their analysis in a monograph. Moreover, some other researchers have developed `R` packages for estimation with RR surveys. Some of them are mentioned as follows. Blair et al. [29] developed the `R` package **rr** to perform regression analyses of sensitive data under some standard RR designs. They also provided tools to conduct power analysis for designing RR items. Heck and Moshagen [141] developed the `R` package **RRreg** to conduct correlation and regression analysis of RR data, simple univariate analysis, bivariate correlations including RR variables, logistic regression with an RR variable, and linear regression with RR variables as predictors. Rueda et al. [142] developed the `R` package **RRTCS** to perform point and interval estimation of linear parameters with data collected from RR surveys under complex sampling designs. Fox et al. [143] extended the existing

implementations by providing generalized regression tools for multiple-group RR designs in the R package **GLMMRR**.

3.11. Non-Randomized Response Techniques

In RR surveys, respondents use a randomization device such as a coin or a deck of cards to generate an outcome that influences the required scrambled answer. However, running a random experiment can be cumbersome and expensive. This has led to the development of NRRTs in recent years. In contrast to RR surveys, in NRR surveys, respondents use an independent non-sensitive question such as their birthday in the questionnaire to obtain their answer to a sensitive question indirectly. In NRR surveys, respondents are expected to give the same response to the questions that are repeated. Some of the more common NRRTs are reviewed below.

3.11.1. Some Common Non-Randomized Response Models

Hidden sensitivity model (HSM): In 2007, Tian et al. [144] proposed a non-randomized HSM to investigate the association between two sensitive binary questions. For example, they considered two variables $X_1, X_2 \in \{0, 1\}$, where $X_1 = 1$ if using drugs and $X_2 = 1$ if having AIDS. This technique is called the HSM because the truthful sensitive attributes of all respondents are hidden. Before Tian et al. [144], for example, Fox and Tracy [65] estimated the correlation between two sensitive questions. Christofides [57] provided an RRT for two sensitive characteristics simultaneously. However, all of these models require the use of randomization devices.

CWM and TRM: In 2008, Yu et al. [28] introduced two NRRTs—the CWM and TRM—for a single sensitive question with binary options. Of which, the CWM can be viewed as a non-randomized version of the original RR model of Warner [25]. However, compared to the original Warner's RR model, the CWM has several advantages, including, e.g., better reproducibility of results and increased cooperation from respondents due to its perceived lower invasiveness. Let X be the sensitive attribute. In these models, X has two categories. For instance, $X \in \{1, 0\}$ with $X = 1$ if having sensitive characteristics and $X = 0$ otherwise. In 2009, Tan et al. [145] showed that the non-randomized TRM has higher relative efficiency and better degree of privacy protection compared to the Warner's RR model [25]. In 2020, Hoffmann et al. [146] conducted a study to compare directly the validity of the CWM and TRM and contrast their performance with a conventional DQ approach.

Multi-category response model (MCRM): In 2009, Tang et al. [147] developed a non-randomized MCRM for surveys with a single categorical sensitive question. This model is suitable for the case of the sensitive variable X with k categories: $X \in \{1, 2, \ldots, k\}$, $k \geq 2$. For example, let $X \in \{1, 2, 3\}$ with $X = 1$ if having never violated traffic laws; $X = 2$ if having ever violated traffic laws once or twice; and $X = 3$ if having violated traffic laws three or more times. A requirement for this model is that at least one value of X, say $X = 1$, is non-sensitive.

Diagonal model (DM): In 2014, Groenitz [148] proposed a survey technique, called a DM, for multi-categorical sensitive variables. The DM is an NRR method to avoid using any randomization device and, hence, reduce the complexity and costs of surveys. That at least one category of the sensitive variable be non-sensitive is not required in the DM. Consequently, one can even apply the DM to attributes, such as income, which are sensitive as a whole.

Parallel model (PM): In 2014, Tian [149] introduced another NRRT, the PM that is a non-randomized version of the randomized unrelated-question model. He explored the asymptotic properties of the ML estimator and its modified version for the proportion of interest. Theoretical comparisons have shown that the PM is generally more efficient than the CWM and TRM for most possible parameter ranges. Additionally, he developed Bayesian methods to analyze survey data gathered from the PM.

By using direct and indirect questions, Perri et al. [150] proposed a procedure to detect the presence of liars in sensitive surveys that allows researchers to evaluate the impact

of untruthful responses on the estimation of the prevalence of a sensitive attribute. They first introduced the theoretical framework, then applied the proposal to the RR method of Warner [25], the unrelated question model [40], the item count technique, the CWM, and the TRM.

3.11.2. Statistical Methods for Non-Randomized Response Models

In 2009, Tian et al. [151] proposed the Bayesian NRR models for surveys including one and two sensitive questions. They derived the exact posterior distributions and their explicit posterior moments, as well as posterior modes via the EM algorithm. They also presented an approach to generate independent and identically-distributed posterior samples for the CWM and TRM, respectively. For the HSM, Tian et al. [144] presented the Bayesian analysis under a conjugate Dirichlet prior as well as some other prior structures. In 2011, Tian et al. [152] developed the formula for determining the sample size required for the non-randomized TRM. This formula was designed to help researchers determine the optimal sample size for a given survey design and level of desired precision.

In 2014, Tang et al. [93] considered a non-randomized TRM to test the equality of the proportions of individuals with a sensitive feature between two independent populations. They derived the Wald, score, and likelihood ratio (LR) tests. They also developed the formulae for determining the sample size. In 2015, Groenitz [119] introduced Bayesian estimation for the DM in [148]. In 2019, Tian et al. [153] developed hidden logistic regression according to the non-randomized PM in Tian [149] to study the relationships between non-sensitive covariates and a sensitive binary response variable. Groenitz [22] developed a general approach for logistic regression analysis with direct data on the covariates and indirect data on the sensitive variable that covers many NRRTs to generate the indirect data. Groenitz [22] derived a general algorithm for the ML estimation and a general procedure for variance estimation.

3.11.3. Real Data with Non-Randomized Response Models

Various applications of NRR designs have appeared in the literature. For instance, Tian et al. [144] described how the non-randomized HSM can be utilized to assess the association between "sex exchange for drugs or money" and "HIV status". Tang et al. [147] illustrated how their NRR method is used to estimate the distribution of the attribute, "number of sex partners", in the population of Korean adolescents. Tang et al. [93] applied a TRM to conduct a simple questionnaire survey to test whether the proportions of college students who had homosexual experience were equal for men and women. The equality of the proportions of college students who had homosexual experience for males and females were examined by the Wald, score, and LR tests. Hoffmann et al. [146] conducted an experimental comparison of the CWM and TRM.

Hoffmann et al. [146] conducted a study on Xenophobia and opposition to reception of refugees in Germany. In a paper-pencil survey of 1,382 students, they estimated prevalence of the two sensitive features, xenophobia and rejection of further refugee admissions, and one non-sensitive control trait with a known prevalence (the first letter of respondents' surnames). They showed that NRRTs provide more valid prevalence estimates for socially undesirable characteristics compared to conventional DQ. The CWM was particularly able to successfully control for the influence of social desirability bias, and outperformed the TRM, presumably because of the favorable influence of the response symmetry found in the CWM but not the TRM. They also found that the sensitivity of two questions was contingent on respondents' political orientation, and that the CWM provided the most valid estimates for respondents for whom these questions were most sensitive. According to these results, they recommended the use of the CWM over the TRM or DQ for highly sensitive topics in a survey's target population. Recently, Chang et al. [2] and Lee et al. [55] studied the experience of one-night stands among freshmen at Feng Chia University in Taiwan in 2016. They used an NRR design via the concept of Warner's RR model [25] and Christofides GRR model [54], respectively. Groenitz [22] re-presented real data on the sales of gas stations in

Germany with the sensitive characteristic sales (with categories low, medium and high) to demonstrate the applicability of the developed general framework. Perri et al. [150] used the CWM and the TM to collect the data and to investigate the problem of racism among students at the University of Calabria, Italy, in 2016, and the phenomenon of workplace mobbing. They showed the estimates for the prevalence of the sensitive attributes under study and evaluated the impact of the liars on the reliability of the final results.

3.11.4. Some Extensions of the Non-Randomized Response Models

Extended crosswise model (ECRM): In the CRM, the sample is not split into multiple groups. Heck et al. [154] introduced the ECRM, where respondents are randomly assigned to two groups. The ECRM not only guarantees the same statistical efficiency as the CRM but also can enable researchers to detect respondents' non-compliance with instructions.

Dual NRR model and alternating NRR model: Wu and Tang [155] proposed the dual NRRT and the alternating NRRT to actively account for deception in the TRM. In the former, the sample is split into two groups, with two different non-sensitive questions. In the latter, although the sample is also split into two groups, only one non-sensitive question is used. Both the two methods have been argued to provide more accurate estimates than the TRM.

Cheating detection triangular model (CDTM): To improve upon the previous IQTs, Meisters et al. [156] proposed the new CDTM. Similar to the cheating detection model of Clark and Desharnais [157], it includes a mechanism for detecting instruction non-adherence and, similar to the TRM, it utilizes simplified instructions to improve respondents' understanding of the procedure. Based on their results, the CDTM appears to be the best choice among the investigated IQTs.

4. Conclusions

We have systematically reviewed the RRT-related works, from the pioneering work of Warner (1965) [25] to the present, according to their respective aspects and to the best of our knowledge. It includes several developments in RR designs as well as statistical methods used in the problems of interest in this field. In each respect, instead of introducing all related works, we re-introduced typical and pioneering works. A more complete view of the evolution of the RRT can be found in the monographs listed in the References.

Author Contributions: Ideas, S.-M.L. and C.-S.L.; writing—original draft preparation, T.-N.L. and P.-L.T.; writing—review and editing, C.-S.L. and S.-M.L.; supervision, S.-M.L. and C.-S.L. All authors have read and agreed to the published version of the manuscript.

Funding: Lee's research was supported by Ministry of Science and Technology (MOST) Grant of Taiwan, ROC, MOST-109-2118-M-035-002-MY3.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1. Expectation and Variance of $\widehat{\theta}_W$ in the Model of Warner [25]

Let λ denote the probability of answering "Yes". Then, $\lambda = P(Yes) = P(A)P(Yes|A) + P(\overline{A})P(Yes|\overline{A}) = \theta p + (1-\theta)(1-p)$. Let n_1 be the number of individuals responding "Yes". n_1 then follows the binomial distribution with parameters n and λ. Its expectation and variance are given by

$$\begin{aligned} \mathrm{E}(n_1) &= n(\theta p + (1-\theta)(1-p)), \\ \mathrm{Var}(n_1) &= n(\theta p + (1-\theta)(1-p))[1 - (\theta p + (1-\theta)(1-p))] \\ &= n\Big(-\theta^2(2p-1)^2 + \theta(2p-1)^2 + p(1-p)\Big). \end{aligned}$$

$\widehat{\theta}_W = \frac{p-1}{2p-1} + \frac{n_1}{(2p-1)n}$ is then an unbiased estimator of θ because

$$\begin{aligned}
\mathrm{E}(\widehat{\theta}_W) &= \frac{p-1}{2p-1} + \frac{\mathrm{E}(n_1)}{(2p-1)n} \\
&= \frac{p-1}{2p-1} + \frac{n(\theta p + (1-\theta)(1-p))}{(2p-1)n} \\
&= \frac{p-1}{2p-1} + \frac{1-p+\theta(2p-1)}{2p-1} \\
&= \theta.
\end{aligned}$$

Moreover, we can get

$$\begin{aligned}
\mathrm{Var}(\widehat{\theta}_W) &= \frac{\mathrm{Var}(n_1)}{(2p-1)^2 n^2} \\
&= \frac{n\big(-\theta^2(2p-1)^2 + \theta(2p-1)^2 + p(1-p)\big)}{(2p-1)^2 n^2} \\
&= \frac{-\theta^2(2p-1)^2 + \theta(2p-1)^2 + p(1-p)}{(2p-1)^2 n} \\
&= \frac{\theta(1-\theta)}{n} + \frac{1}{4n}\left[\frac{1}{(2p-1)^2} - 1\right] \\
&= \frac{\theta(1-\theta)}{n} + \frac{p(1-p)}{n(2p-1)^2}.
\end{aligned}$$

Appendix A.2. Expectation and Variance of $\widehat{\theta}_C$ in the Model of Christofides [54]

Now, each sampled person is provided with a randomization device that is used to generate the integers $1, 2, \ldots, L$ with probabilities $P_1, P_2, \ldots, P_L$, respectively. Using the randomization device, the individual generates one of these L numbers and reports how far the generated number is away from $L+1$ if she/he had the sensitive characteristic or from 0 otherwise.

Let Y_i take on the value $L+1$ if individual i had the sensitive characteristic and the value 0 if not. Clearly $P(Y_i = L+1) = \theta$ and $P(Y_i = 0) = 1-\theta$. Let T_i be the integer produced by individual i using the randomization device. The reported number is then $D_i = |Y_i - T_i|$ whose pmf is given by

$$P(D_i = d) = (1-\theta)P_d + \theta P_{L+1-d}, \quad d = 1, 2, \ldots, L.$$

Direct calculation shows that

$$\begin{aligned}
\mathrm{E}(D_i) &= \sum_{d=1}^{L} dP(D_i = d) \\
&= \sum_{d=1}^{L} d[(1-\theta)P_d + \theta P_{L+1-d}] \\
&= \sum_{d=1}^{L} dP_d + \theta \sum_{d=1}^{L} d(P_{L+1-d} - P_d).
\end{aligned}$$

Because

$$\begin{aligned}\sum_{d=1}^{L} d(P_{L+1-d} - P_d) &= (P_L - P_1) + 2(P_{L-1} - P_2) + \cdots + L(P_1 - P_L)\\ &= (L-1)P_1 + (L-3)P_2 + (L-5)P_3 + \cdots + (L-(2L-1))P_L\\ &= (L+1-2)P_1 + (L+1-4)P_2 + (L+1-6)P_3 + \cdots + (L+1-(2L))P_L\\ &= (L+1)\sum_{d=1}^{L} P_d - 2P_1 - 4P_2 - \cdots - 2LP_L\\ &= (L+1)\times 1 - 2(P_1 + 2P_2 + \cdots + LP_L)\\ &= L+1-2\sum_{d=1}^{L} dP_d,\end{aligned}$$

it can yield

$$\begin{aligned}\mathrm{E}(D_i) &= \sum_{d=1}^{L} dP_d + \theta\left(L+1-2\sum_{d=1}^{L} dP_d\right)\\ &= \mathrm{E}(T_i) + \theta(L+1-2\mathrm{E}(T_i)).\end{aligned}$$

Similarly, one can obtain

$$\begin{aligned}\mathrm{E}(D_i^2) &= \sum_{d=1}^{L} d^2 P(D_i = d)\\ &= \sum_{d=1}^{L} d^2((1-\theta)P_d + \theta P_{L+1-d})\\ &= \sum_{d=1}^{L} d^2 P_d + \theta\sum_{d=1}^{L} d^2(P_{L+1-d} - P_d).\end{aligned}$$

We have

$$\begin{aligned}&\sum_{d=1}^{L} d^2(P_{L+1-d} - P_d)\\ &= (P_L - P_1) + 2^2(P_{L-1} - P_2) + \cdots + L^2(P_1 - P_L)\\ &= (L^2-1)P_1 + ((L-1)^2 - 2^2)P_2 + ((L-2)^2 - 3^2)P_3 + \cdots + ((L-(L-1))^2 - L^2)P_L\\ &= (L+1)(L-1)P_1 + (L+1)(L-3)P_2 + (L+1)(L-5)P_3 + \cdots + (L+1)(L-(2L-1))P_L\\ &= (L+1)[(L-1)P_1 + (L-3)P_2 + (L-5)P_3 + \cdots + (L-(2L-1))P_L]\\ &= (L+1)[(L+1-2)P_1 + (L+1-4)P_2 + (L+1-6)P_3 + \cdots + (L+1-2L)P_L]\\ &= (L+1)\left[(L+1)\sum_{d=1}^{L} P_d - 2P_1 - 4P_2 - 6P_3 - \cdots - 2LP_L\right]\\ &= (L+1)\left[L+1-2\sum_{d=1}^{L} dP_d\right].\end{aligned}$$

Thus,

$$\begin{aligned}\mathrm{E}(D_i^2) &= \sum_{d=1}^{L} d^2 P_d + \theta \sum_{d=1}^{L} d^2 (P_{L+1-d} - P_d) \\ &= \sum_{d=1}^{L} d^2 P_d + \theta \left\{ (L+1) \left[L + 1 - 2 \sum_{d=1}^{L} d P_d \right] \right\} \\ &= \mathrm{E}(T_i^2) + \theta \{ (L+1)[L+1-2\mathrm{E}(T_i)] \}.\end{aligned}$$

Accordingly,

$$\begin{aligned}&\mathrm{Var}(D_i) \\ &= \mathrm{E}(D_i^2) - (\mathrm{E}(D_i))^2 \\ &= \mathrm{E}(T_i^2) + \theta \{ (L+1)[L+1-2\mathrm{E}(T_i)] \} - \{ \mathrm{E}(T_i) + \theta [L+1-2\mathrm{E}(T_i)] \}^2 \\ &= \mathrm{Var}(T_i) + \theta \{ (L+1)[L+1-2\mathrm{E}(T_i)] \} - 2\mathrm{E}(T_i)\theta [L+1-2\mathrm{E}(T_i)] - \{ \theta [L+1-2\mathrm{E}(T_i)] \}^2 \\ &= \mathrm{Var}(T_i) + \theta(1-\theta)[L+1-2\mathrm{E}(T_i)]^2.\end{aligned}$$

Let $\overline{D} = \frac{1}{n} \sum_{i=1}^{n} D_i$ and define the estimator

$$\widehat{\theta}_C = \frac{\overline{D} - \mathrm{E}(T_i)}{L+1-2\mathrm{E}(T_i)},$$

provided that $L + 1 - 2\mathrm{E}(T_i) \neq 0$. Then,

$$\begin{aligned}\mathrm{E}(\widehat{\theta}_C) &= \frac{\mathrm{E}(\overline{D}) - \mathrm{E}(T_i)}{L+1-2\mathrm{E}(T_i)} \\ &= \frac{\mathrm{E}(D_i) - \mathrm{E}(T_i)}{L+1-2\mathrm{E}(T_i)} \\ &= \frac{\mathrm{E}(T_i) + \theta[L+1-2\mathrm{E}(T_i)] - \mathrm{E}(T_i)}{L+1-2\mathrm{E}(T_i)} \\ &= \theta,\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}(\widehat{\theta}_C) &= \frac{\mathrm{Var}(\overline{D})}{[L+1-2\mathrm{E}(T_i)]^2} \\ &= \frac{\mathrm{Var}(D_i)}{n[L+1-2\mathrm{E}(T_i)]^2} \\ &= \frac{\mathrm{Var}(T_i) + \theta(1-\theta)[L+1-2E(T_i)]^2}{n[L+1-2\mathrm{E}(T_i)]^2} \\ &= \frac{\theta(1-\theta)}{n} + \frac{\mathrm{Var}(T_i)}{n[L+1-2E(T_i)]^2}.\end{aligned}$$

References

1. Rueda, M.; Cobo, B.; Perri, P.F. Randomized response estimation in multiple frame surveys. *Int. J. Comput. Math.* **2020**, *97*, 189–206. [CrossRef]
2. Chang, P.C.; Pho, K.H.; Lee, S.M.; Li, C.S. Estimation of parameters of logistic regression for two-stage randomized response technique. *Comput. Stat.* **2021**, *36*, 2111–2133. [CrossRef]
3. Huang, K.C. A Survey technique for estimating the proportion and sensitivity in a dichotomous finite population. *Stat. Neerl.* **2004**, *58*, 75–82. [CrossRef]
4. Mehta, S.; Aggarwal, P. Bayesian estimation of sensitivity level and population proportion of a sensitive characteristic in a binary optional unrelated question RRT model. *Commun. Stat.-Theory Methods* **2018**, *47*, 4021–4028. [CrossRef]

5. Narjis, G.; Shabbir, J. Bayesian analysis of optional unrelated question randomized response models. *Commun. Stat.-Theory Methods* **2021**, *50*, 4203–4215. [CrossRef]
6. Sihm, J.S.; Chhabra, A.; Gupta, S.N. An optional unrelated question RRT model. *Involve* **2016**, *9*, 195–209. [CrossRef]
7. Tourangeau, R.; Yan, T. Sensitive questions in surveys. *Psychol. Bull.* **2007**, *133*, 859–883. [CrossRef]
8. Krumpal, I.; Voss, T. Sensitive questions and trust: Explaining respondents' behavior in randomized response surveys. *SAGE Open* **2020**, 1–17. [CrossRef]
9. Hsieh, S.H.; Lee, S.M.; Tu, S.H. Randomized response techniques for a multi-level attribute using a single sensitive question. *Stat. Pap.* **2018**, *59*, 291–306. [CrossRef]
10. Hsieh, S.H.; Lukusa, M.T.W. Comparison of estimators for multi-level randomized response data: Evidence from a case of sexual identity. *Field Methods* **2021**, *33*, 85–103. [CrossRef]
11. Hsieh, S.H.; Tu, S.H.; Lee, S.M.; Wang, C.W. Application of the randomized response technique in the 2012 presidential election of Taiwan. *Surv. Res.-Method Appl.* **2016**, *35*, 81–109.
12. Hsieh, S.H.; Lee, S.M.; Li, C.S. A two-stage multilevel randomized response technique with proportional odds models and missing covariates. *Sociol. Methods Res.* **2022**, *51*, 439–467. [CrossRef]
13. Hsieh, S.H.; Lee, S.M.; Li, C.S.; Tu, S.H. An alternative to unrelated randomized response techniques with logistic regression analysis. *Stat. Method. Appl.* **2016**, *25*, 601–621. [CrossRef] [PubMed]
14. Lee, S.M.; Peng, T.C.; Tapsoba, J.D.D.; Hsieh, S.H. Improved estimation methods for unrelated question randomized response techniques. *Commun. Stat.-Theory Methods* **2017**, *46*, 8101–8112. [CrossRef]
15. Hyman, H. Do they tell the truth? *Public Opin. Q.* **1944**, *8*, 557–559. [CrossRef]
16. Tourangeau, R.; Rips, L.J.; Rasinski, K. *The Psychology of Survey Response,* 1st ed.; Cambridge University Press: Cambridge, UK, 2000.
17. Kerkvliet, J. Cheating by economics students: A comparison of survey results. *J. Econ. Educ.* **1994**, *25*, 121–133. [CrossRef]
18. Preisendörfer, P.; Wolter, F. Who is telling the truth? A validation study on determinants of response behavior in surveys. *Public Opin. Q.* **2014**, *78*, 126–146. [CrossRef]
19. van der Heijden, P.G.M.; van Gils, G.; Bouts, J.A.N.; Hox, J.J. A comparison of randomized response, computer-assisted self-interview, and face-to-face direct questioning: Eliciting sensitive information in the context of welfare and unemployment benefit. *Sociol. Methods Res.* **2000**, *28*, 505–537. [CrossRef]
20. Hsieh, S.H.; Perri, P.F. Estimating the proportion of non-heterosexuals in Taiwan using Christofides' randomized response model: A comparison of different estimation methods. *Soc. Sci. Res.* **2021**, *93*, 102475. [CrossRef]
21. Chaudhuri, A.; Christofides, T.C. *Indirect Questioning in Sample Surveys;* Springer: Berlin/Heidelberg, Germany, 2013.
22. Groenitz, H. Logistic regression analyses for indirect data. *Commun. Stat.-Theory Methods* **2018**, *47*, 3838–3856. [CrossRef]
23. Ibbett, H.; Jones, J.P.; St. John, F.A. Asking sensitive questions in conservation using randomised response techniques. *Biol. Conserv.* **2021**, *260*, 109191. [CrossRef] [PubMed]
24. Nuno, A.; John, F.A.S. How to ask sensitive questions in conservation: A review of specialized questioning techniques. *Biol. Conserv.* **2015**, *189*, 5–15. [CrossRef]
25. Warner, S.L. Randomized response: A survey technique for eliminating evasive answer bias. *J. Am. Stat. Assoc.* **1965**, *60*, 63–69. [CrossRef] [PubMed]
26. Dalton, D.R.; James, C.W.; Catherine, M.D. Using the unmatched count technique (UCT) to estimate base rates for sensitive behavior. *Pers. Psychol.* **1994**, *47*, 817–827. [CrossRef]
27. Droitcour, J.; Caspar, R.A.; Hubbard, M.L.; Parsley, T.L.; Visscher, W.; Ezzati, T.M. The item count technique as a method of indirect questioning: A review of its development and a case study application. In *Measurement Errors in Surveys*; Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., Sudman, S., Eds.; Wiley: New York, NY, USA, 1991; pp. 185–210.
28. Yu, J.W.; Tian, G.L.; Tang, M.L. Two new models for survey sampling with sensitive characteristic: Design and analysis. *Metrika* **2008**, *67*, 251–263. [CrossRef]
29. Blair, G.; Imai, K.; Zhou, Y.Y. Design and analysis of the randomized response technique. *J. Am. Stat. Assoc.* **2015**, *110*, 1304–1319. [CrossRef]
30. Sungkawichai, T.; Thongsata, P.; Paka, T.; Laoharenoo, A.; Vatiwutipong, P. Forced randomized response protocol using arbitrary random variable. *Curr. Appl. Sci. Technol.* **2023**, *23*, 1–10. [CrossRef]
31. Tian, G.L.; Tang, M.L. *Incomplete Categorical Data Design: Non-Randomized Response Techniques for Sensitive Questions in Surveys;* Chapman & Hall/CRC: Boca Raton, FL, USA, 2013.
32. Arnab, R. *Survey Sampling Theory and Applications;* Academic Press: Cambridge, MA, USA, 2017.
33. Chaudhuri, A.; Mukerjee, R. *Randomized Response: Theory and Techniques;* CRC Press: New York, NY, USA, 1988.
34. Chaudhuri, A. *Randomized Response and Indirect Questioning Techniques in Surveys;* Chapman & Hall/CRC: Boca Raton, FL, USA, 2011.
35. Chaudhuri, A.; Christofides, T.C.; Rao, C.R. Data Gathering, analysis and protection of privacy through randomized response techniques: Qualitative and quantitative human traits. In *Handbook of Statistics 34*; Chaudhuri, A., Christofides, T.C., Rao, C.R., Eds.; Elsevier: Amsterdam, The Netherlands, 2016; pp. 29–41.
36. Fox, J.A. *Randomized Response and Related Methods: Surveying Sensitive Data;* Sage Publication: Thousand Oaks, CA, USA, 2016.

37. Tracy, D.S.; Mangat, N.S. Some developments in randomized response sampling during the last decade—A follow up of review by Chauduri and Mukerjee. *J. Appl. Stat. Sci.* **1996**, *4*, 147–158.
38. Lensvelt-Mulders, G.J.L.M.; Van Der Heijden, P.G.M.; Laudy, O.; Van Gils, G. A validation of a computer-assisted randomized response survey to estimate the prevalence of fraud in social security. *J. R. Stat. Soc. Ser. A* **2006**, *169*, 305–318. [CrossRef]
39. Horvitz, D.G.; Shah, B.V.; Simmons, W.R. The unrelated question randomized response model. *Proc. Soc. Stat. Sect. Am. Stat. Assoc.* **1967**, *62*, 65–72.
40. Greenberg, B.G.; Abul-Ela, A.L.A.; Simmons, W.R.; Horvitz, D.G. The unrelated question randomized response model: Theoretical framework. *J. Am. Stat. Assoc.* **1969**, *64*, 520–539. [CrossRef]
41. Chaudhuri, A.; Mukerjee, R. Optionally randomized response techniques. *Calcutta Stat. Assoc. Bull.* **1985**, *34*, 225–230. [CrossRef]
42. Bhargava, M.; Singh, R. A modified randomization device for Warner's model. *Statistica* **2000**, *60*, 315–322.
43. Kim, J.M.; Warde, W.D. A stratified Warner's randomized response model. *J. Stat. Plan. Inference* **2004**, *120*, 155–165. [CrossRef]
44. Abbasi, A.M.; Shad, M.Y.; Ahmed, A. On partial randomized response model using ranked set sampling. *PLoS ONE* **2022**, *17*, e0277497. [CrossRef]
45. Zapata, Z.; Sedory, S.A.; Singh, S. An innovative improvement in Warner's randomized response device for evasive answer bias. *J. Stat. Comput. Simul.* **2023**, *93*, 298–311. [CrossRef]
46. Edgell, S.E.; Duchan, K.L.; Himmelfarb, S. An empirical test of the unrelated question randomized response technique. *Bull. Psychon. Soc.* **1992**, *30*, 153–156. [CrossRef]
47. Shaw, P.; Chaudhuri, A. Further improvements on unrelated characteristic models in randomized response techniques. *Commun. Stat.-Theory Methods* **2022**, *51*, 7305–7321. [CrossRef]
48. Chaudhuri, A.; Shaw, P. Generating randomized response by inverse Bernoullian trials in unrelated characteristics model. *Model Assist. Stat. Appl.* **2016**, *11*, 235–245.
49. Mangat, N.S.; Singh, R. An alternative randomized response procedure. *Biometrika* **1990**, *77*, 439–442. [CrossRef]
50. Mangat, N.S. An improved randomized response strategy. *J. R. Stat. Soc. Ser. B-Stat. Methodol.* **1994**, *56*, 93–95. [CrossRef]
51. Chang, H.J.; Liang, D.H. A two-stage unrelated randomized response procedure. *Aust. N. Z. J. Stat.* **1996**, *38*, 43–51.
52. Gjestvang, C.R.; Singh, S. A new randomized response model. *J. R. Stat. Soc. Ser. B-Stat. Methodol.* **2006**, *68*, 523–530. [CrossRef]
53. Vishwakarma, G.K.; Kumar, A.; Kumar, N. Two-stage unrelated randomized response model to estimate the prevalence of a sensitive attribute. *Comput. Stat.* **2023**, 1–26. [CrossRef]
54. Christofides, T.C. A generalized randomized response technique. *Metrika* **2003**, *57*, 195–200. [CrossRef]
55. Lee, S.M.; Le, T.N.; Tran, P.L.; Li, C.S. Investigating the association of a sensitive attribute with a random variable using the Christofides generalised randomised response design and Bayesian methods. *J. R. Stat. Soc. Ser. C* **2022**, *71*, 1471–1502. [CrossRef]
56. Chaudhuri, A. Christofides' randomized response technique in complex sample surveys. *Metrika* **2004**, *60*, 223–228. [CrossRef]
57. Christofides, T.C. Randomized response technique for two sensitive characteristics at the same time. *Metrika* **2005**, *62*, 53–63. [CrossRef]
58. Christofides, T.C. Randomized response in stratified sampling. *J. Stat. Plan. Infer.* **2005**, *128*, 303–310. [CrossRef]
59. Lee, C.S.; Sedory, S.A.; Singh, S. Estimating at least seven measures of qualitative variables from a single sample using randomized response technique. *Stat. Probab. Lett.* **2013**, *83*, 399–409. [CrossRef]
60. Perri, P.F.; Pelle, E.; Stranges, M. Estimating induced abortion and foreign irregular presence using the randomized response crossed model. *Soc. Indic. Res.* **2016**, *129*, 601–618. [CrossRef]
61. Abul-Ela, A.L.A.; Greenberg, G.G.; Horvitz, D.G. A multi-proportions randomized response model. *J. Am. Stat. Assoc.* **1967**, *62*, 990–1008. [CrossRef]
62. Liu, P.T.; Chow, L.P. The efficiency of the multiple trial randomized response technique. *Biometrika* **1976**, *32*, 607–618. [CrossRef]
63. Barksdale, W.B. New Randomized Response Techniques for Control of Non-Sampling Errors in Surveys. Ph.D. Dissertation, University of North Carolina, Chapel Hill, NC, USA, 1971. [CrossRef]
64. Drane, W. On the theory of randomized responses to two sensitive questions. *Commun. Stat.-Theory Methods* **1976**, *5*, 565–574. [CrossRef]
65. Fox, J.A.; Tracy, P.E. Measuring associations with randomized response. *Soc. Sci. Res.* **1984**, *13*, 188–197.
66. Ewemooje, O.S. Estimating two sensitive characters with equal probabilities of protection. *Cogent Math.* **2017**, *4*, 1319607. [CrossRef]
67. Ewemooje, O.S.; Amahia, G.N. Improved randomized response technique for two sensitive attributes. *Afr. Stat.* **2015**, *10*, 839–852. [CrossRef]
68. Ewemooje, O.S.; Amahia, G.N. Improving the efficiency of randomized response technique for two sensitive characters. *FUTA J. Res. Sci.* **2016**, *12*, 65–72. [CrossRef]
69. Batool, F.; Shabbir, J. A two-stage design for multivariate estimation of proportions. *Commun. Stat.-Theory Methods* **2016**, *45*, 5412–5426.
70. Xu, T.; Sedory, S.A.; Singh, S. Two sensitive characteristics and their overlap with two questions per card. *Biom. J.* **2021**, *63*, 1688–1705.
71. Chung, R.S.W.; Chu, A.M.Y.; So, M.K.P. Bayesian randomized response technique with multiple sensitive attributes: The case of information systems resource misuse. *Ann. Appl. Stat.* **2018**, *12*, 1969–1992. [CrossRef]

72. Chu, A.M.Y.; Omori, Y.; So, H.Y.; So, M.K.P. A multivariate randomized response model for sensitive binary data. *Econom. Stat.* **2022**, 1–20. [CrossRef]
73. Greenberg, B.G.; Kuebler, R.R., Jr.; Abernathy, J.R.; Horvitz, D.G. Application of the randomized response technique in obtaining quantitative data. *J. Am. Stat. Assoc.* **1971**, *66*, 243–250. [CrossRef]
74. Gupta, S.; Gupta, B.; Singh, S. Estimation of sensitivity level of personal interview survey questions. *J. Stat. Plan. Inference* **2002**, *100*, 239–247. [CrossRef]
75. Grewal, I.S.; Bansal, M.L.; Singh, S. Estimation of population mean of a stigmatized quantitative variable using double sampling. *Statistica* **2003**, *63*, 79–88. [CrossRef]
76. Hussain, Z.; Shabbir, J. Estimation of mean of a sensitive quantitative variable. *J. Stat. Res.* **2007**, *41*, 83–92. [CrossRef]
77. Hussain, Z.; Shakeel, S.; Cheema, S.A. Estimation of stigmatized population total: A new additive quantitative randomized response model. *Commun. Stat.-Theory Methods* **2022**, *51*, 8741–8753.
78. Gupta, S.; Zhang, J.; Khalil, S.; Sapra, P. Mitigating lack of trust in quantitative randomized response technique models. *Commun. Stat.-Simul. Comput.* **2022**, 1–9.
79. Warner, S.L. The linear randomized response model. *J. Am. Stat. Assoc.* **1971**, *366*, 884–888. [CrossRef]
80. Goodstadt, M.S.; Gruson, V. The randomized response technique: A test on drug use. *J. Am. Stat. Assoc.* **1975**, *70*, 814–818. [CrossRef]
81. Maddala, G.S. *Limited-Dependent and Qualitative Variables in Econometrics*; Cambridge University Press: Cambridge, UK, 1983.
82. Ewemooje, O.S.; Adeniyi, I.O.; Adediran, A.A.; Molefe, W.B.; Adebola, F.B. Bayesian estimation in alternative tripartite randomized response techniques. *Sci. Afr.* **2023**, *19*, e01584. [CrossRef]
83. Abernathy, J.R.; Greenberg, B.G.; Horvitz, D.G. Estimates of induced abortion in urban North Carolina. *Demography* **1970**, *7*, 19–29. [CrossRef]
84. Winkler, R.L.; Franklin, L.A. Warner's randomized response model: A Bayesian approach. *J. Am. Stat. Assoc.* **1979**, *74*, 207–214. [CrossRef]
85. Hsieh, S.H.; Lee, S.M.; Shen, P.S. Semiparametric analysis of randomized response data with missing covariates in logistic regression. *Comput. Stat. Data Anal.* **2009**, *53*, 2673–2692. [CrossRef]
86. Hsieh, S.H.; Lee S.M.; Shen, P.S. Logistic regression analysis of randomized response data with missing covariates. *J. Stat. Plan. Inference* **2010**, *140*, 927–940. [CrossRef]
87. van der Heijden, P.G.M.; van Gils, G. Some logistic regression models for randomized response data. In *Statistical Modelling, Proceedings of the 11th International Workshop on Statistical Modelling, Orvieto, Italy, 15–19 July 1996*; Forcina, A., Marchetti, G.M., Hatzinger, R., Falmacci, G., Eds.; Graphos: Città di Castello, Italy, 1996; pp. 341–348.
88. van den Hout, A.; van der Heijden, P.G.M.; Gilchrist, R. The logistic regression model with response variables subject to randomized response. *Comput. Stat. Data Anal.* **2007**, *51*, 6060–6069. [CrossRef]
89. Krumpal, I. Estimating the prevalence of xenophobia and anti-Semitism in Germany: A comparison of randomized response and direct questioning. *Soc. Sci. Res.* **2012**, *41*, 1387–1403. [CrossRef]
90. Ostapczuk, M.; Musch, J.; Mashagen, M. A randomized-response investigation of the education effect in attitudes towards foreigners. *Eur. J. Soc. Psychol.* **2009**, *39*, 920–931. [CrossRef]
91. Arnab, R.; Mothupi, T. Randomized response techniques: A case study of the risky behaviors' of students of a certain University. *Model Assist. Stat. Appl.* **2015**, *10*, 421–430. [CrossRef]
92. Rueda, M.M.; Cobo, B.; López-Torrecillas, F. Measuring inappropriate sexual behavior among university students: Using the randomized response technique to enhance self-reporting. *Sex. Abus.* **2020**, *32*, 320–334. [CrossRef] [PubMed]
93. Tang, M.L.; Wu, Q.; Tian, G.L.; Guo, J.H. Two-sample non randomized response techniques for sensitive questions. *Commun. Stat.-Theory Methods* **2014**, *43*, 408–425. [CrossRef]
94. Scheers, N.J.; Dayton, C.M. Covariate randomized response models. *J. Am. Stat. Assoc.* **1988**, *83*, 969–974. [CrossRef]
95. Hsieh, S.H.; Perri, P.F. A logistic regression extension for the randomized response simple and crossed models: Theoretical results and empirical evidence. *Sociol. Methods Res.* **2022**, *51*, 1244–1281. [CrossRef]
96. Mieth, L.; Mayer, M.M.; Hoffmann, A.; Buchner, A.; Bell, R. Do they really wash their hands? Prevalence estimates for personal hygiene behaviour during the COVID-19 pandemic based on indirect questions. *BMC Public Health* **2021**, *21*, 12. [CrossRef]
97. Reiber, F.; Bryce, D.; Ulrich, R. Self-protecting responses in randomized response designs: A survey on intimate partner violence during the coronavirus disease 2019 pandemic. *Sociol. Methods Res.* **2022**, 1–32. [CrossRef]
98. Striegel, H.; Ulrich, R.; Simon, P. Randomized response estimates for doping and illicit drug use in elite athletes. *Drug Alcohol Depend.* **2010**, *106*, 230–232. [CrossRef]
99. Christiansen, A.V.; Frenger, M.; Chirico, A.; Pitsch, W. Recreational athletes' use of performance-enhancing substances: Results from the first European randomized response technique survey. *Sport. Med.-Open* **2023**, *9*, 1. [CrossRef]
100. Mielecka-Kubień, Z.; Toniszewski, M. Estimation of illicit drug use among high school students in the Silesian voivodship (Poland) with the use of the randomized response technique. *Math. Popul. Stud.* **2022**, *29*, 47–57. [CrossRef]

101. Liu, P.T.; Chow, L.P. A new discrete quantitative randomized response model. *J. Am. Stat. Assoc.* **1976**, *71*, 72–73. [CrossRef]
102. Burgstaller, L.; Feld, L.P.; Pfeil, K. Working in the shadow: Survey techniques for measuring and explaining undeclared work. *J. Econ. Behav. Organ.* **2022**, *200*, 661–671. [CrossRef]
103. Bar-Lev, S.K.; Bobovich, E.; Boukai, B. A common conjugate prior structure for several randomized response models. *Test* **2003**, *12*, 101–113. [CrossRef]
104. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B-Stat. Methodol.* **1977**, *39*, 1–22.
105. Bourke, P.D.; Moran, M.A. Estimating proportions from randomized response data using the EM algorithm. *J. Am. Stat. Assoc.* **1988**, *83*, 964–968. [CrossRef]
106. Lee, S.M.; Tran, P.L.; Le, T.N.; Li, C.S. Prediction of a sensitive feature under indirect questioning via Warner's randomized response technique and latent class model. *Mathematics* **2023**, *11*, 345. [CrossRef]
107. van den Hout, A.; Kooiman, P. Estimating the linear regression model with categorical covariates subject to randomized response. *Comput. Stat. Data Anal.* **2006**, *50*, 3311–3323. [CrossRef]
108. Nandram, B.; Yu, Y. Bayesian analysis of sparse counts obtained from the unrelated question design. *Int. J. Stat. Probab.* **2019**, *8*, 66–84. [CrossRef]
109. Hussain, Z.; Shabbir, J.; Riaz, M. Bayesian estimation using Warner's randomized response model through simple and mixture prior distributions. *Commun. Stat.-Simul. Comput.* **2011**, *40*, 147–164. [CrossRef]
110. Migon, H.S.; Tachibana, V.M. Bayesian approximations in randomized response model. *Comput. Stat. Data Anal.* **1997**, *24*, 401–409. [CrossRef]
111. Pitz, G.F. Bayesian analysis of random response models. *Psychol. Bull.* **1980**, *87*, 209–212. [CrossRef]
112. Fidler, D.S.; Kleinknecht, R.E. Randomized response versus direct questioning: Two data-collection methods for sensitive information. *Psychol. Bull.* **1977**, *84*, 1045–1049. [CrossRef]
113. O'Hagan, A. Bayes linear estimators for randomized response models. *J. Am. Stat. Assoc.* **1987**, *82*, 580–585. [CrossRef]
114. Oh, M.S. Bayesian analysis of randomized response models: A Gibbs sampling approach. *J. Korean Stat. Soc.* **1994**, *23*, 463–482.
115. Unnikrishnan, N.K.; Kunte, S. Bayesian analysis for randomized response models. *Sankhyā Indian J. Stat. (Ser. B)* **1999**, *61*, 422–432.
116. Hussain, Z.; Shabbir, J. Bayesian estimation of population proportion in Kim and Warde mixed randomized response technique. *Electron. J. Appl. Stat. Anal.* **2012**, *5*, 213–225.
117. Song, J.J.; Kim, J.M. Bayesian analysis of randomized response sum score variables. *Commun. Stat.-Theory Methods* **2012**, *41*, 1875–1884. [CrossRef]
118. Adepetun, A.O.; Adewara, A.A. Bayesian analysis of Kim and Warde randomized response technique using alternative priors. *Am. J. Comput. Appl. Math.* **2014**, *4*, 130–140.
119. Groenitz, H. Using prior information in privacy-protecting survey designs for categorical sensitive variables. *Stat. Pap.* **2015**, *56*, 167–189. [CrossRef]
120. Song, J.J.; Kim, J.M. Bayesian estimation of rare sensitive attribute. *Commun. Stat.-Simul. Comput.* **2017**, *64*, 4154–4160. [CrossRef]
121. Kerkvliet, J. Estimating a logit model with randomized data: The case of cocaine use. *Aust. N. Z. J. Stat.* **1994**, *36*, 9–20. [CrossRef]
122. Boruch, R.F. Assuring confidentiality of responses in social research: A note on strategies. *Am. Sociol.* **1971**, *6*, 308–311.
123. Kuk, A.Y.C. Asking sensitive questions indirectly. *Biometrika* **1990**, *77*, 436–438. [CrossRef]
124. Cruyff, M.J.; Böckenholt, U.; Van Der Heijden, P.G.; Frank, L.E. A review of regression procedures for randomized response data, including univariate and multivariate logistic regression, the proportional odds model and item response model, and self-protective responses. In *Handbook of Statistics 34*; Chaudhuri, A., Christofides, T.C., Rao, C.R., Eds.; Elsevier: Amsterdam, The Netherlands, 2016; pp. 287–315.
125. Ronning, G. Randomized response and the binary probit model. *Econ. Lett.* **2005**, *86*, 221–228. [CrossRef]
126. Rubin, D.B. Inference and missing data. *Biometrika* **1976**, *63*, 581–592. [CrossRef]
127. Horvitz, D.G.; Thompson, D.J. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* **1952**, *47*, 663–685. [CrossRef]
128. Lee, S.M.; Lukusa, T.M.; Li, C.S. Estimation of a zero-inflated Poisson regression model with missing covariates via nonparametric multiple imputation methods. *Comput. Stat.* **2020**, *35*, 725–754. [CrossRef]
129. Stoklosa, J.; Lee, S.M.; Hwang, W.H. Closed population capture–recapture models with measurement error and missing observations in covariates. *Stat. Sin.* **2019**, *29*, 589–610. [CrossRef]
130. Wang, D.; Chen, S.X. Empirical likelihood for estimating equations with missing values. *Ann. Stat.* **2009**, *37*, 490–517. [CrossRef]
131. Rubin, D.B. Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*; American Statistical Association: Alexandria, VA, USA, 1978; Volume 1, pp. 20–34.
132. Fay, R.E. Alternative paradigms for the analysis of imputed survey data. *J. Am. Stat. Assoc.* **1996**, *91*, 490–498. [CrossRef]
133. Lee, S.M.; Le, T.N.; Tran, P.L.; Li, C.S. Estimation of logistic regression with covariates missing separately or simultaneously via multiple imputation methods. *Comput. Stat.* **2022**, 1–35. [CrossRef]

134. Spiegelhalter, D.J.; Best, N.; Carlin, B.; van der Linde, A. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B-Stat. Methodol.* **2002**, *64*, 583–639. [CrossRef]
135. Chib, S. Marginal likelihood from the Gibbs output. *J. Am. Stat. Assoc.* **1995**, *90*, 1313–1321. [CrossRef]
136. Wentland, J.J.; Reissing, E. Casual sexual relationships: Identifying definitions for one night stands, booty calls, fuck buddies, and friends with benefits. *Can. J. Hum. Sex.* **2014**, *23*, 167–177. [CrossRef]
137. Kaspar, K.; Buß, L.V.; Rogner, J.; Gnambs, T. Engagement in one-night stands in Germany and Spain: Does personality matter? *Pers. Individ. Differ.* **2016**, *92*, 74–79. [CrossRef]
138. Hox, J.; Lensvelt-Mulders, G. Randomized response analysis in Mplus. *Struct. Equ. Model.* **2004**, *11*, 615–620. [CrossRef]
139. Sehra, S. Two-Stage Optional Randomized Response Models. Master's Thesis, The University of North Carolina, Greensboro, NC, USA, 2008.
140. Jann, B. RRLOGIT: Stata Module to Estimate Logistic Regression for rAndomized Response Data. 2011. Available at Research Papers in Economics (RePEc). Available online: https://ideas.repec.org/c/boc/bocode/s456203.html (accessed on 12 May 2011).
141. Heck, D.W.; Moshagen, M. RRreg: An R package for correlation and regression analyses of randomized response data. *J. Stat. Softw.* **2018**, *85*, 1–29. [CrossRef]
142. Rueda, M.D.M.; Cobo, B.; Arcos, A. RRTCS: An R package for randomized response techniques in complex surveys. *Appl. Psychol. Meas.* **2016**, *40*, 78–80. [CrossRef]
143. Fox, J.P.; Klotzke, K.; Veen, D. Generalized linear randomized response modeling using GLMMRR. *arXiv* **2021**, arXiv:2106.10171.
144. Tian, G.L.; Yu, J.W.; Tang, M.L.; Geng, Z. A new non-randomized model for analyzing sensitive questions with binary outcomes. *Statist. Med.* **2007**, *26*, 4238–4252. [CrossRef]
145. Tan, M.T.; Tian, G.L.; Tang, M.L. Sample surveys with sensitive questions: A nonrandomized response approach. *Am. Stat.* **2009**, *63*, 9–16. [CrossRef]
146. Hoffmann, A.; Meisters, J.; Musch, J. On the validity of non-randomized response techniques: An experimental comparison of the crosswise model and the triangular model. *Behav. Res. Methods* **2020**, *52*, 1768–1782. [CrossRef] [PubMed]
147. Tang, M.L.; Tian, G.L.; Tang, N.S.; Liu, Z.Q. A new non-randomized multicategory response model for surveys with a single sensitive question: Design and analysis. *J. Kor. Statist. Soc.* **2009**, *38*, 339–349. [CrossRef]
148. Groenitz, H. A new privacy-protecting survey design for multichotomous sensitive variables. *Metrika* **2014**, *77*, 211–224. [CrossRef]
149. Tian, G.L. A new non-randomized response model: The parallel model. *Stat. Neerl.* **2014**, *68*, 293–323. [CrossRef]
150. Perri, P.F.; Manoli, E.; Christofides, T.C. Assessing the effectiveness of indirect questioning techniques by detecting liars. *Stat. Pap.* **2022**, 1–24. [CrossRef]
151. Tian, G.L.; Yuen, K.C.; Tang, M.L.; Tan, M.T. Bayesian non-randomized response models for survey with sensitive questions. *Stat. Interface* **2009**, *2*, 13–25.
152. Tian, G.L.; Tang, M.L.; Liu, Z.; Tan, M.; Tang, N.S. Sample size determination for the non-randomized triangular model for sensitive questions in a survey. *Statist. Meth. Med. Res.* **2011**, *20*, 159–173. [CrossRef]
153. Tian, G.L.; Liu, Y.; Tang, M.L. Logistic regression analysis of non-randomized response data collected by the parallel model in sensitive surveys. *Aust. N. Z. J. Stat.* **2019**, *61*, 134–151. [CrossRef]
154. Heck, D.W.; Hoffmann, A.; Moshagen, M. Detecting nonadherence without loss in efficiency: A simple extension of the crosswise model. *Behav. Res. Methods* **2018**, *50*, 1895–1905. [CrossRef]
155. Wu, Q.; Tang, M.L. Non-randomized response model for sensitive survey with noncompliance. *Stat. Methods Med. Res.* **2016**, *25*, 2827–2839. [CrossRef] [PubMed]
156. Meisters, J.; Hoffmann, A.; Musch, J. A new approach to detecting cheating in sensitive surveys: The cheating detection triangular model. *Sociol. Methods Res.* **2022**, 1–31. [CrossRef]
157. Clark, S.J.; Desharnais, R.A. Honest answers to embarrassing questions: Detecting cheating in the randomized response model. *Psychol. Methods* **1998**, *3*, 160–168. [CrossRef]