RESEARCH REPORT SERIES
(Statistics #2020-06)

A Review of Rigorous Randomized Response Methods for Protecting Respondent's Privacy and Data Confidentiality

Tapan K. Nayak[1, 2]

[1]Center for Statistical Research and Methodology, U.S. Census Bureau
[2] Department of Statistics, George Washington University

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: October 6, 2020

A Review of Rigorous Randomized Response Methods for Protecting Respondent's Privacy and Data Confidentiality *

Tapan K. Nayak†

Abstract

Randomized response (RR) methods for protecting respondent's privacy when collecting data on sensitive characteristics have been proposed and discussed for over fifty years. The basic ideas of RR have also been used to develop the post-randomization method (PRAM) for protecting data confidentiality. Both RR and PRAM randomize true responses using specified probabilities and the choice of those probabilities is central to designing RR methods and PRAM. However, most papers do not give clear guidance on how to choose the transition probabilities. Some rigorous approaches have appeared only recently. This paper reviews the essential elements of RR and PRAM, some important differences between the two, and designing RR methods and PRAM for achieving certain precise privacy and confidentiality protection goals. In particular, we discuss (i) designing an RR survey to guarantee that a randomized response would not reveal much information about the respondent, in a precise sense, and (ii) devising PRAM to strictly control identification risks when releasing microdata.

Key words and Phrases: Data utility; identity disclosure; minimaxity; post-randomization; privacy criteria; transition probability.

*The views expressed in this article are those of the author and not those of the U.S. Census Bureau.

†Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC 20233 and Department of Statistics, George Washington University, Washington, DC 20052, email:tapan@gwu.edu.

1. Introduction

The primary objective of randomized response (RR) methods is to protect respondent's privacy and thereby reduce false responses when collecting data on sensitive or stigmatizing characteristics, such as tax evasion, drug use, gambling and abortion. The first RR method, introduced by Warner (1965), concerns interview surveys of a binary characteristic, where the population consists of a sensitive group A (e.g., drug users) and its complement A^c (e.g., not drug users). In Warner's method, a respondent answers one of the two questions: Q_1: Do you belong to A? and Q_2: Do you belong to A^c? Each respondent selects a question by performing a specified random experiment, unobserved by the interviewer, with a given device, e.g., a spinner or a shuffled deck of cards bearing the two questions. The respondent answers the selected question. Thus, the interviewer does not know the question that a respondent answers. The experiment imposes specific and known probabilities, say p and $1-p$, of selecting Q_1 and Q_2. The value of p determines both the degree of privacy protection and the amount of statistical information loss due to randomization. Thus, to design Warner's method, one should first select p and then construct an experiment for implementing it.

Following Warner's (1965) pioneering work, numerous other RR methods with different mechanisms for randomizing the true responses have been proposed for both categorical and quantitative variables. We refer interested readers to the books Chaudhuri and Mukerjee (1988), Chaudhuri (2010) and Chaudhuri et al. (2016) for discussion of various methods and further references. In this article, we consider RR methods only for categorical variables. An RR method transforms the true responses probabilistically. For a true response, the RR output is generated from a predetermined probability distribution on an output space. In Section 2, we briefly review a general framework and some key theoretical results that are needed for later sections. RR methods for quantitative variables are structurally quite different and can generally be viewed as (additive or multiplicative) random noise infusion to the true values.

For many years, research on RR methods was done primarily by statisticians and for use

in face-to-face interview surveys, which requires experiments that are easy to understand and perform correctly and randomization devices that are simple and portable. Various randomization devices and mechanisms have been proposed as different RR methods. Recently, Blair et al. (2015) stated that "our extensive search yields only a handful of published studies that use the randomized response method to answer substantive questions." Evidently, the theoretical advances in RR methods have not been used much in real surveys. Perhaps, one reason for that gap is the increasing adoption of mail, telephone and online surveys, replacing interview surveys, in the past several decades. Notably, Holbrook and Krosnick (2010) tested RR methods in one telephone and eight Internet surveys of American adults and found that respondents were either unable or unwilling to implement the RR mechanisms properly.

Interestingly, starting around the beginning of this century, interest in RR methods has grown tremendously among computer scientists and in new dimensions. The explosive growth of automated data capture by companies from business transactions, online searches, postings and other activities have raised much public awareness and concerns about privacy. Computer scientists have taken substantial interest in developing theory and methods for privacy-preserving data mining and data publishing; see e.g., Aggarwal and Yu (2008), Chen et al. (2009) and Fung et al. (2019). New privacy concepts and measures have been developed and RR methods have been implemented by leading companies such as Google, Apple and Microsoft; see Erlingsson et al. (2014), Ding et al. (2017) and Cormode et al. (2018). Perhaps, one impetus for this resurgence and expansion of interest in RR techniques is that in online data capturing, the actual randomization can be carried out easily and accurately by computer programs. Consequently, recent research on RR theory and methods has focused appropriately on the choice of the randomization probabilities, leaving aside ancillary features of possible physical mechanisms for implementing them.

Gouweleeuw et al. (1998) used the basic ideas of RR to introduce the post-randomization method (PRAM) for perturbing categorical data to protect data confidentiality. It also randomizes the true responses, similar to RR. But, the data agency performs randomization after data

collection and so, the randomization probabilities may be chosen based on the data set. We discuss some important differences between RR and PRAM in Section 2. For both methods, the choice of the transition probabilities is critical. Intuitively, one should try to choose those to minimize data utility (or statistical information) loss while meeting privacy and confidentiality protection goals. Some precise and practical privacy and confidentiality protection goals have been proposed and investigated only recently. One primary objective of this article is to review some rigorous approaches to privacy and confidentiality protection via RR and PRAM. We also attempt to cover some important works that appeared in computer science literature.

In Section 2, we describe the essential elements of RR and PRAM and some basic mathematical results. We do not aim to give a comprehensive review of RR and PRAM, but attempt to highlight some important points that we think have not been appreciated well. In Section 3, we discuss a rigorous notion of privacy protection. The basic idea is that an RR procedure should guarantee that an output would not reveal much new information about the respondent's true value. A formal development of this idea compares prior and posterior probabilities. This criterion is closely connected to *local differential privacy*, which has received considerable attention in recent years. We also discuss a necessary and sufficient condition that an RR method must satisfy in order to guarantee such privacy. In Section 4, we discuss optimal RR methods for providing specified privacy. In particular, we describe an admissiblity result under a very general view of data utility, and optimal methods under certain criteria. Section 5 relates to data confidentiality protection. Specifically, we consider identity disclosure control in microdata release and review a procedure that uses PRAM to guarantee an upper bound for the probability of correctly identifying any unit in perturbed microdata. Section 6 is devoted some concluding remarks.

2. Basic Elements of RR and PRAM

In this section, we describe the mathematical structures of RR and PRAM. Applying RR to multiple variables is equivalent to applying it to the cross-classification of those variables. Let X denote a categorical variable or cross-classification of several variables that we want to subject to RR. Let $\mathcal{S}_X = \{c_1, \ldots, c_k\}$ denote the set of possible categories of X. Let $\pi_i = P(X = c_i), i = 1, \ldots, k$, and $\pi = (\pi_1, \ldots, \pi_k)'$, which is unknown. The goal of an RR survey is to obtain information about π while protecting respondent's privacy. The basic idea of RR is to collect a stochastic transformation of each respondent's true category. Let Z denote the output variable with output space $\mathcal{S}_Z = \{d_1, \ldots, d_m\}$. A special case is $\mathcal{S}_Z = \mathcal{S}_X$. In general, $\mathcal{S}_Z$ may be different from $\mathcal{S}_X$, with possibly $m \neq k$. For each input, the output is selected according to some probabilities specified by the RR method. Let $p_{ij} = P(Z = d_i | X = c_j), i = 1, \ldots, m, j = 1, \ldots, k$, denote the transition probabilities of an RR method. Obviously, $\sum_i p_{ij} = 1$ for $j = 1, \ldots, k$. The transition probability matrix (TPM), $P = ((p_{ij}))$, is chosen during the planning of an RR method.

For any given P, one can devise multiple physical experiments to implement it (see Nayak et al., 2016). Under the common assumption of truthful respondent participation, all statistical properties of an RR method depend solely on its TPM. The choice of the experiment for implementing a given P does not affect the mathematical properties of the method. Thus, formal assessments and comparisons of RR methods should be made via their TPMs, ignoring the randomization experiments. However, as Leysieffer and Warner (1976), Fligner et al. (1977), Nayak (1994) and others have noted, when comparing different RR methods some papers have incorrectly matched disparate features of the experiments and drawn false conclusions. We consider P as the *design* of an RR method and discuss the planning and analysis of an RR method in terms of P.

A common assumption for analyzing RR data is multinomial sampling, i.e., random sampling from an infinite population or simple random sampling with replacement if the population is

finite. Let $\lambda_i = P(Z = d_i), i = 1, \ldots, m$, and $\lambda = (\lambda_1, \ldots \lambda_m)'$. Also, let S_i denote the frequency of $Z = d_i$ and $\mathbf{S} = (S_1, \ldots, S_k)'$. Then, $\mathbf{S}$ is multinomially distributed, $\mathbf{S} \sim Mult(n, \lambda)$, where

$$\lambda = P\pi \tag{2.1}$$

and n is the sample size. We can use $\mathbf{S}$ to make inferences about λ. If $m = k$ and P is nonsingular, from an estimator of λ one can obtain an estimator of π via (2.1). In particular, $\hat{\lambda} = \mathbf{S}/n$, which is the MLE (and UMVUE) of λ, yields $\hat{\pi} = P^{-1}\hat{\lambda} = P^{-1}(\mathbf{S}/n)$. It can be seen that $\hat{\pi}$ is an unbiased estimator of π and

$$Var(\hat{\pi}) = \frac{1}{n}(D_\pi - \pi\pi') + \frac{1}{n}[P^{-1}D_\lambda(P^{-1})' - D_\pi], \tag{2.2}$$

where D_π is a diagonal matrix with diagonal elements $\pi_1, \ldots, \pi_k$ and D_λ is defined similarly (see Chaudhuri and Mukerjee, 1988, p. 43). The first term on the right side of (2.2) is the sampling variance and the last term is the additional variance due to randomization.

If $m < k$ or $\text{rank}(P) < k$, then the model for $\mathbf{S}$ is not identifiable with respect to π and hence π is not estimable from RR data. Thus, for estimability of π, one should only consider $m \geq k$ and $\text{rank}(P) = k$. While $m = k$ is quite common, several methods with $m > k$ have also been proposed, e.g., Leysieffer and Warner (1976), Kuk (1990) and Christofides (2003). Also, we shall see in Section 4.2 that $m > k$ in some optimal designs. We should mention that while most authors discussed estimation of π under multinomial sampling, Padmawar and Vijayan (2000), Chaudhuri (2001, 2004) and Nayak and Adeshiyan (2009) derived estimators of π under general sampling designs.

Now we turn our attention to PRAM. Like RR, it randomizes the true responses with known probabilities. However, that is done after data collection and by data agencies. These two practical matters yield some important differences between RR and PRAM. Here, we state some special and helpful features of PRAM, which we believe have not been well recognized. First, in

PRAM, the transition probabilities may be chosen based on the entire data set, containing all true responses (which is not possible in RR surveys as the responses are randomized during data collection). Indeed, unbiased PRAM, discussed below, requires the TPM to depend on the data. When the TPM is data dependent, it is a random matrix and mathematical results in RR for fixed P, e.g., (2.2), may not hold true. Second, randomization may be applied selectively only to the responses with high disclosure risks. Note that agencies remove all direct identifiers, such as name and address, before releasing data. So, a respondent's values (true or randomized) are not revealed directly. In contrast, in RR surveys, a respondent's identity is known to the data collector. Third, related to the previous point, one may partition the data into homogeneous sets and then apply PRAM separately within the partition sets with possibly different TPMs. One method that utilizes data partitioning and unbiased PRAM is described in Section 5. Fourth, the randomization is carried out by a computer program, without needing a physical experiment. Fifth, in PRAM $\mathcal{S}_Z = \mathcal{S}_X$ and thus $m = k$ and the original and perturbed data appear in the same format. Sixth, the transition probabilities may not be known publicly. In particular, the transition probabilities cannot be published conveniently or helpfully when those are chosen diversely using the observed data, as in Section 5. There, the data agency should use a carefully designed unbiased PRAM to well preserve data utility, so that the released data may practically be treated as original data for making inferences.

Next, to describe *unbiased* PRAM, let T_i and S_i denote the frequency of c_i in the original and perturbed data, respectively, and let $\mathbf{T} = (T_1, \ldots, T_k)'$ and $\mathbf{S} = (S_1, \ldots, S_k)'$. When the data are collected by multinomial sampling, a PRAM with TPM P is said to be unbiased (Gouweleeuw et al. (1998) called this invariant) if

$$P\mathbf{T} = \mathbf{T}. \tag{2.3}$$

It can be easily verified that the solution space of (2.3) is a convex set and a trivial solution is $P = I$. Gouweleeuw et al. (1998) gave two methods for finding nontrivial solutions.

Unbiased PRAM was motivated by the fact that (2.3) implies $E[\mathbf{S}|\mathbf{T}] = P\mathbf{T}$ and hence

$\hat{\pi}_* = \mathbf{S}/n$ is an unbiased estimator of π. Also, $\hat{\pi}_*$ is always a probability vector and it can be calculated without using P or its inverse. Thus, π can be estimated easily from perturbed data. Nayak and Adeshiyan (2016) derived and explored the exact variance of $\hat{\pi}_*$. In particular, they gave a decomposition of the variance into sampling variance and added variance due to PRAM, similar to (2.2). They also discussed estimation of π under a general sampling plan and unbiased PRAM. There, P is called unbiased if $P\hat{\pi} = \hat{\pi}$, where $\hat{\pi}$ is an appropriately weighted (and usually unbiased) estimator of π based on the original data.

We refer interested readers to in Gouweleeuw et al. (1998), Willenborg and De Waal (2001), Van den Hout and Van der Heijedn (2002), Van den Hout and Elamir (2006) and Shlomo and Skinner (2010) for additional discussion and applications of PRAM. The main task of designing RR and PRAM is choosing the transition probabilities. Naturally, suitable choices should depend on privacy and confidentiality protection goals. In the following sections we discuss some recently developed precise privacy and confidentiality protection goals and methods for achieving those goals.

3. Privacy Protection by RR

Most privacy measures in statistics literature were developed for the situation where the survey variable is binary with one sensitive category and the response is also binary, see e.g., Leysieffer and Warner (1976), Lanke (1976), Fligner et al. (1977), Nayak (1994) and Zhimin and Zaizai (2012). Using common terminology, let A and A^c denote the two categories of X, of which A is sensitive, and let Y (for yes) and N (for no) denote the two response categories. The privacy measures in the binary case are mostly functions of the two posterior probabilities $P(A|Y)$ and $P(A|N)$, where the prior probabilities of A and A^c are their population proportions, say π_A and $1 - \pi_A$. For example, Lanke's (1976) measure of the degree of privacy protection is $\max\{P(A|Y), P(A|N)\}$. This and most other measures for the binary case depend also on π_A, which is unknown. A common suggestion is: choose the RR design parameters $P(Y|A)$ and

$P(Y|A^c)$ such that a chosen privacy measure does not exceed a threshold at some π_A. Here, the designer of the survey selects the privacy measure, the threshold and π_A. For many privacy measures this is equivalent to requiring

$$\max\left\{\frac{P(Y|A)}{P(Y|A^c)}, \frac{P(Y|A^c)}{P(Y|A)}\right\} \leq \gamma, \tag{3.1}$$

where γ is determined by the privacy measure, privacy threshold and π_A.

3.1. Strict Privacy Criteria

The preceding approaches do not consider an intruder's personal knowledge about respondents. The following approach, recently initiated by computer scientists, focuses on the basic goal of privacy protection, which is limiting the amount of information an intruder might gain about a respondent from his/her randomized response. Informally, the main idea, due to Evfimievski et al. (2003), is that we should view a privacy breach as an intruder gaining much new information about a respondent and an RR design should guarantee that no such privacy breaches would occur. Furthermore, an intruder's information (or opinion) should be expressed precisely using subjective probability.

Formally, let $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ denote an intruder's (subjective) prior distribution for a respondent's true value of X, and for any $Q \subseteq \mathcal{S}_X$, let $P_\alpha(Q)$ and $P_\alpha(Q|d_i)$ denote, respectively, the intruder's prior and posterior probabilities of $\{X \in Q\}$, given $Z = d_i$. Here, Q represents a "property" of the respondent. Clearly,

$$P_\alpha(Q) = \sum_{j:c_j \in Q} \alpha_j \quad \text{and} \quad P_\alpha(Q|d_i) = \sum_{j:c_j \in Q} P_\alpha(X = c_j|Z = d_i),$$

where

$$P_\alpha(X = c_j|Z = d_i) = \frac{P_\alpha(X = c_j, Z = d_i)}{P_\alpha(Z = d_i)} = \frac{\alpha_j p_{ij}}{\sum_{l=1}^k \alpha_l p_{il}}.$$

One idea is that to protect respondent's privacy, we should guarantee that $P_\alpha(Q|d_i)$ would never

be much higher or lower than $P_\alpha(Q)$, i.e., $P_\alpha(Q|d_i)$ would always be "close" to $P_\alpha(Q)$. This idea yields different privacy criteria for different specifications of desired closeness between prior and posterior probabilities.

Evfimievski et al. (2003) used a specific "closeness" criterion and introduced ρ_1-to-ρ_2 privacy as follows.

Definition 3.1. *Let $0 < \rho_1 < \rho_2 < 1$ be two numbers. (a) An RR procedure is said to permit an upward ρ_1-to-ρ_2 privacy breach with respect to $Q \subseteq \mathcal{S}_X$ and a prior distribution α if*

$$P_\alpha(Q) < \rho_1 \quad \textit{and} \quad P_\alpha(Q|d_i) > \rho_2$$

for some $1 \leq i \leq m$ with $P_\alpha(Z = d_i) > 0$. Similarly, a procedure permits a downward ρ_2-to-ρ_1 privacy breach with respect to Q and α if $P_\alpha(Q) > \rho_2$ and $P_\alpha(Q|d_i) < \rho_1$ for some d_i with $P_\alpha(Z = d_i) > 0$.

(b) An RR procedure provides ρ_1-to-ρ_2 privacy protection if it does not permit an upward ρ_1-to-ρ_2 or a downward ρ_2-to-ρ_1 privacy breach with respect to any Q and α.

Using the ratio of posterior to prior probabilities as a measure of closeness between the two, Nayak et al. (2015) defined the following.

Definition 3.2. *For a given $\beta > 1$, an RR procedure admits a β-factor privacy breach, with respect to $Q \subseteq \mathcal{S}_X$ and a prior α if $P_\alpha(Q) > 0$ and*

$$\frac{P_\alpha(Q|d_i)}{P_\alpha(Q)} > \beta \quad \textit{or} \quad \frac{P_\alpha(Q|d_i)}{P_\alpha(Q)} < \frac{1}{\beta}$$

for some d_i such that $P_\alpha(Z = d_i) > 0$. An RR procedure provides β-factor privacy if it does not allow a β-factor breach with respect to any Q and α.

Chai and Nayak (2018) developed and explored the preceding ideas generally and we review their main results below. A general criterion for considering two probabilities as sufficiently

"close" gives (explicitly or implicitly) for each $0 < p < 1$, an interval $[l_p, u_p]$ that consists of all values that are considered sufficiently close to p (taking a closed interval for simplicity). So, if a prior probability is p, a privacy breach occurs if and only if a corresponding posterior probability falls outside the interval $[l_p, u_p]$. This yields two functions $h_l(p) \equiv l_p$ and $h_u(p) \equiv u_p$, which specify the lower and upper breach boundaries. Thus, a general criterion may be viewed as a pair of given functions $h_l(p)$ and $h_u(p)$. Considering this, Chai and Nayak (2018) introduced the following.

Definition 3.3. *Let h_l and h_u be two functions from $[0,1]$ to $[0,1]$ such that $0 \leq h_l(a) \leq a \leq h_u(a) \leq 1$ for all $0 \leq a \leq 1$. An RR procedure is said to satisfy privacy with respect to h_l and h_u if*

$$h_l(P_\alpha(Q)) \leq P_\alpha(Q|d_i) \leq h_u(P_\alpha(Q)) \tag{3.2}$$

for all $\alpha, Q \subseteq \mathcal{S}_X$ and $i = 1, \ldots, m$.

This definition says that a posterior probability p_* is sufficiently close to the corresponding prior p if $h_l(p) \leq p_* \leq h_u(p)$. Geometrically, a (prior, posterior) pair (p, p_*) is a point in the unit square, of which the regions below h_l and above h_u constitute the privacy breach region (PBR) of the criterion in Definition 3.3. A privacy satisfying RR method must not yield any prior-posterior pair that falls in the PBR. Conversely, the PBR of an RR procedure P is the collection of all non-attainable (prior, posterior) pairs under P. The privacy holding region of P or the complement of the PBR (with respect to the unit square) is $\{(p, p_*), 0 \leq p, p_* \leq 1 : P_\alpha(Q) = p$ and $P_\alpha(Q|d_i) = p_*$ for some d_i, α and $Q \subseteq \mathcal{S}_X\}$.

3.2. Privacy Characterization

A natural question is: for given h_l and h_u, how to find a TPM that satisfies (3.2)? The first $\leq$ in (3.2) is equivalent to $P_\alpha(Q^c|d_i)) \leq 1 - h_l(1 - P_\alpha(Q^c))$. Considering this for all $Q \subseteq \mathcal{S}_X$, and defining $h(a) = \min\{h_u(a), 1 - h_l(1-a)\}$, for $0 \leq a \leq 1$, it follows that an RR procedure

P satisfies (3.2) if and only if

$$P_\alpha(Q|d_i) \leq h(P_\alpha(Q)) \tag{3.3}$$

for all $i = 1, \ldots, m$ and all α and $Q \subseteq \mathcal{S}_X$ such that $0 < P_\alpha(Q) < 1$. Chai and Nayak (2018) gave a necessary and sufficient condition for satisfying (3.3) using the following concept.

Definition 3.4. *(Nayak et al., 2015). The ith row parity of P is defined as*

$$\eta_i(P) = \max\left\{\frac{p_{ij}}{p_{il}} \;\middle|\; j, l = 1, \ldots, k\right\} = \frac{\max_j\{p_{ij}\}}{\min_j\{p_{ij}\}},$$

with the convention $0/0 = 1$ and $a/0 = \infty$ for any $a > 0$. The parity of P is defined as $\eta(P) = \max_i\{\eta_i(P)\}$.

Theorem 3.1. *For a given h, an RR procedure P satisfies (3.3) if and only if $\eta(P) \leq B(h)$, where*

$$B(h) = \inf_{0<p<1}\Big(\frac{1-p}{p}\Big)\Big(\frac{h(p)}{1-h(p)}\Big)$$

and $h(p)/[1-h(p)] = \infty$ when $h(p) = 1$.

The necessary and sufficient condition in Theorem 3.1 depends on h only through $B(h)$ and on P only through its parity $\eta(P)$. Thus, $B(h)$ quantifies the privacy demand of h and $\eta(P)$ quantifies the privacy level of P. These two measures are useful for comparing privacy demands of different PBRs and privacy levels of various RR procedures, respectively. Chai and Nayak (2018) also showed that an RR procedure P with $\eta(P) = \gamma > 1$ guarantees (3.3) for all h such that

$$h(p) \geq h_{(\gamma)}(p) \equiv \frac{\gamma p}{1 + (\gamma - 1)p}, \quad 0 < p < 1. \tag{3.4}$$

Thus, $h_{(\gamma)}(.)$ in (3.4) is the precise upper breach boundary of any P with parity γ. Choosing h_l and h_u in practical applications may appear a difficult task. But, the preceding discussions give helpful guidance. Specifically, we only need to consider the functions in $\mathcal{H} = \{h_{(\gamma)}(.); \gamma > 1\}$ and choose one of those for the upper breach boundary. The precise lower breach boundary

corresponding to $h_{(\gamma)}(.)$ is $\tilde{h}_{(\gamma)}(p) = 1 - h_{(\gamma)}(1-p)$. In practice, the plots of the precise PBRs for various γ might be helpful in selecting an appropriate PBR.

Theorem 3.1 says that in order to satisfy the criterion in Definition 3.3, each row parity of P must not exceed an upper bound, determined by h_l and h_u. The privacy condition (3.1) in the binary case with one sensitive category is similar. It gives an upper bound for the parity of just one row, corresponding to the response Y. We should mention that under (3.1), Nayak (1994) showed that the transition probabilities of an optimal design are $P(Y|A) = 1, P(Y|A^c) = 1/\gamma$ and hence $P(N|A) = 0$ and $P(N|A^c) = 1 - 1/\gamma$. However, this optimal design asks all respondents in the sensitive group to answer 'Yes,' which might be uncomfortable for some respondents. It also implies $P(A|N) = 0$, which might encourage some respondents to give the innocuous answer 'No.' Thus, the (mathematically) optimal design may not be suitable in real surveys. This indicates that (3.1) is inadequate and we should impose additional restrictions on the transition probabilities.

We also want to mention the following criterion that has received considerable attention in recent years, especially from computer scientists, see e.g., Kairouz et al. (2016), Wang et al. (2016), Duchi et. al. (2018) and Ye and Barg (2018).

Definition 3.5. *An RR design provides ϵ-local differential privacy (ϵ-LDP), for $\epsilon > 0$, if*

$$\sup_{Q \subseteq \mathcal{S}_Z} \sup_{c_i, c_j \in \mathcal{S}_X} \frac{P(Z \in Q | X = c_i)}{P(Z \in Q | X = c_j)} \leq e^{\epsilon}.$$

Chai and Nayak (2018) proved that an RR procedure provides ϵ-LDP if and only if

$$\frac{P_\alpha(Q)}{1 + (\gamma - 1)(1 - P_\alpha(Q))} \leq P_\alpha(Q|d_i) \leq \frac{\gamma P_\alpha(Q)}{1 + (\gamma - 1)P_\alpha(Q)}$$

for all α, Q and d_i, where $\gamma = e^{\epsilon}$. This shows that ϵ-LDP coincides with Definition 3.3, with $h_l(a) = a/[1 + (\gamma - 1)(1 - a)]$ and $h_u(a) = \gamma a/[1 + (\gamma - 1)a]$. This describes the PBR of ϵ-LPD, which may be used to communicate its privacy promises in terms of bounds on an intruder's

possible information gain. It also follows that an RR design P provides ϵ-LDP if and only if $\eta(P) \leq \gamma = e^{\epsilon}$.

4. Comparison of RR Designs

We have seen that to satisfy the privacy requirement of either ϵ-LDP or Definition 3.3 we must use a design with $\eta(P) \leq \gamma$, for a given value of γ. For any given $\gamma > 1$, let $\mathcal{C}_\gamma = \{P_{m\times k} : \eta(P) \leq \gamma\}$, which is the class of all privacy preserving designs at level γ. Typically, $\mathcal{C}_\gamma$ is large and a natural question is: how should we choose a design from $\mathcal{C}_\gamma$ for practical application? As we noted earlier, for estimability of π, we should choose $P_{m\times k}$ with $m \geq k$ and full rank. Also, as Chai and Nayak (2018) discussed, P should not have any proportional rows, to be concise. Two designs are statistically equivalent if one can be obtained by merging the proportional rows of the other one. Intuitively, we should choose a P from $\mathcal{C}_\gamma$ that satisfies the preceding two conditions and maximizes data utility. However, data utility is a complex matter and it may be assessed in different ways. In the following we first review an admissibility result under a broad view of data utility and then discuss design selection under certain optimality criteria.

4.1. Admissible Designs

Blackwell (1951, 1953) introduced a general criterion for comparing experiments, which in our context says the following.

Definition 4.1. *An RR design $P_{m\times k}$ is said to be sufficient for (or at least as informative as) another RR design $A_{r\times k}$, to be denoted $P \succeq A$, if there exists a transition probability matrix $C_{r\times m}$ such that $A = CP$.*

If $P \succeq A$ and also $A \succeq P$, then A and P are equivalent, and P is better than A if $P \succeq A$ but $A \not\succeq P$. Furthermore, P is said to be admissible if there does not exist another design that is better than P.

If $P \succeq A$, then applying A is equivalent to randomizing the true responses first using P and

then randomizing the outputs of P using C. Intuitively, P should be more informative than A because the second randomization with C inflicts additional loss of statistical information. Formally, $P \succeq A$ implies that given any loss function and any inference rule δ based on the data from A, there exists a rule δ_* based on P whose risk function is no larger than the risk function of δ. In this sense, if $P \succeq A$, then P is universally at least as good as A. Chai and Nayak (2018) proved that two designs $P_{m\times k}$ and $A_{r\times k}$ in $\mathcal{C}_\gamma$ are equivalent if and only if $m = r$ and $A = CP$, where C is a permutation matrix, i.e., A can be obtained by permuting the rows of P, or just reordering the elements of the output space. Logically, we should use only admissible designs. Here, an important question is: how do we know if a given design P is admissible or not? The following result of Chai and Nayak (2018) answers this question.

Theorem 4.1. *For any given γ, an RR design $P \in \mathcal{C}_\gamma$ is admissible if and only if (i) $\eta_i(P) = \gamma$ for all i (i.e., each row parity is γ) and (ii) each row of P contains exactly two distinct values.*

Now, we discuss an important RR method, viz. the RAPPOR algorithm, proposed recently by Erlingsson et al. (2014). Google, Apple, Microsoft and other companies have been using it for online data capture; see, Ding et al. (2017) and Cormode et al. (2018). The basic method applies ϵ-LDP and works as follows. It represents all true responses with indicator vectors $X = (X_1, \ldots, X_k)$. Specifically, if the true response c_i, then the ith component of X is 1 and all other components are 0. RAPPOR's randomization changes each component of $(X_1, \ldots, X_k)$ independently with probability $p = 1/(\sqrt{\gamma}+1)$ and produces an output vector $Z = (Z_1, \ldots, Z_k)$.

The output space of RAPPOR is $\mathcal{S}_Z = \{z = (z_1, \ldots, z_k) : z_i = 0 \text{ or } 1 \text{ for } i = 1, \ldots, k\}$, which contains 2^k elements. So, RAPPOR's TPM is of order $2^k \times k$. The transition probabilities can be calculated easily. Let $x^{(i)}$ denote the indicator vector for true response c_i, i.e., $x^{(i)} = (x_1, \ldots, x_k)$, where $x_i = 1$ and $x_j = 0$ for all $j \neq i$. For any $z \in \mathcal{S}_z$, let $t_z = \sum_j z_j$. Then, it can be seen that

$$P((z_1, \ldots, z_k)|x^{(i)}) = \begin{cases} p^{t_z-1}(1-p)^{k-t_z+1}, & \text{if } z_i = 1 \\ p^{t_z+1}(1-p)^{k-t_z-1}, & \text{if } z_i = 0. \end{cases} \tag{4.1}$$

Let $\vec{1} = (1, \ldots, 1)$ and $\vec{0} = (0, \ldots, 0)$. From (4.1), we see that $P(\vec{1}|x^{(i)}) = p^{k-1}(1-p)$ and $P(\vec{0}|x^{(i)}) = p(1-p)^{k-1}$ for all i. So, the two rows of the RAPPOR's TPM, corresponding to $z = \vec{0}$ and $z = \vec{1}$, have parity 1 (not γ). This shows, in view of Theorem 4.1, the RAPPOR's design is not admissible.

We should mention that all other rows of RAPPOR's TPM satisfy the conditions of Theorem 4.1. So, the RAPPOR algorithm can be modified to make it admissible. Specifically, removing the two rows corresponding to $\vec{0}$ and $\vec{1}$ and normalizing the remaining matrix gives an admissible design. RAPPOR also gives a method for estimating π from perturbed data. The RAPPOR estimator is unbiased but not efficient. Chai and Nayak (2019) derived a better unbiased estimator that is also minimax under certain conditions.

4.2. Comparison of RR Designs

For $k = 2$, i.e., binary X, it follows from Theorem 4.1 that essentially only one design, given below, is admissible and hence it is the best design.

Theorem 4.2. *For binary X, an optimal RR design in $\mathcal{C}_\gamma$ is $P_{2\times 2}$ with $p_{11} = p_{22} = \gamma(\gamma+1)^{-1}$ and $p_{12} = p_{21} = (\gamma+1)^{-1}$.*

The optimal design in Theorem 4.2 is a Warner's design. For $k \geq 3$, many designs are admissible and choosing an optimal design requires additional criteria. In the following, we review some recent results.

Agrawal et al. (2009) presented the following optimality result in a special case. They required ρ_1-to-ρ_2 privacy. An RR design P satisfies ρ_1-to-ρ_2 privacy if and only if $\eta(P) \leq \gamma$, with $\gamma = [\rho_2(1-\rho_1)]/[\rho_1(1-\rho_2)]$. They considered the special case of $\mathcal{S}_Z = \mathcal{S}_X$. This implies that $m = k$. Additionally, they considered only symmetric P. Under these conditions, they proposed to take any P with the minimum condition number as an optimal design. The condition number of a symmetric positive definite matrix is defined as the ratio of its largest and smallest eigenvalues. To justify the criterion, they stated that the stability of numerical calculations

with a matrix decreases as its condition number increases. They were mainly concerned with computing the inverse of P, which is often used for calculating an estimate of π and its variance. They proved that among all $P \in \mathcal{C}_\gamma$ that are also symmetric, the matrix P_0 with elements $p_{ii} = \gamma/(\gamma + k - 1), i = 1, \ldots, k$, and $p_{ij} = 1/(\gamma + k - 1)$ for $i \neq j$ has the minimum condition number. Thus, P_0 is the best design by their criterion.

Chai and Nayak (2018) also considered the special case of $\mathcal{S}_Z = \mathcal{S}_X$. Thus, the true values are randomized within the categories X. Here, the diagonal elements of P are the probabilities of keeping the true responses unchanged. Intuitively, we should change the true responses as little as possible to minimize data utility loss. This suggests to use a design $P \in \mathcal{C}_\gamma$ that has large diagonal values. One measure of "largeness" of the diagonal values of P is $\sum_i p_{ii}$, the trace of P. With this, a design $P \in \mathcal{C}_\gamma$ with the largest trace may be considered a best design at privacy level γ. Chai and Nayak (2018) proved that the optimal design P_0 of Agrawal et al. (2009), given above, is also the best design under the maximum trace criterion. Note that unlike Agrawal et al. (2009), this approach does not require P to be symmetric, although the optimal design P_0 is so.

Next, we review a minimax approach, recently investigated by Chai and Nayak (2019). Assuming multinomial sampling and squared error loss, they considered optimum determination of an RR *strategy*, which consists of a design P and an estimator $\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_k)$ of π. Under squared error loss, the risk function of an RR strategy $(P, \hat{\pi})$ is

$$R(P, \hat{\pi}; \pi) = E_{P,\pi}\Big[\|\hat{\pi} - \pi\|^2\Big] = E_{P,\pi}\Big[\sum_{i=1}^{k}(\hat{\pi}_i - \pi_i)^2\Big], \tag{4.2}$$

where the expectation is with respect to both sampling and randomization. They considered unbiased estimators of π that are also linear in $\mathbf{S}$, i.e., $\hat{\pi} = L\mathbf{S}/n$ for some matrix L, where $\mathbf{S}$ is the frequency vector of $d_1, \ldots, d_m$ and the divisor n is used for mathematical simplicity. As, $E(\mathbf{S}) = nP\pi$, $\hat{\pi} = L\mathbf{S}/n$ is unbiased if and only if $LP = I$. Recall that privacy protection requires that $P \in \mathcal{C}_\gamma$, for specified γ. Then, an RR strategy $(P_*, \tilde{\pi} = L_*S/n)$ is minimax among

all privacy satisfying strategies if $P_* \in \mathcal{C}_\gamma, P_* L_* = I$ and

$$\sup_{\pi} E_{P_*,\pi}\Big[\|\frac{L_* S}{n} - \pi\|^2\Big] = \inf_{P\in\mathcal{C}_\gamma} \inf_{L:LP=I} \sup_{\pi} E_{P,\pi}\Big[\|\frac{LS}{n} - \pi\|^2\Big].$$

The RR design P_* is a minimax design.

To describe the minimax strategy, derived in Chai and Nayak (2019), for given $\gamma > 1$ and $k \geq 2$, define

$$f(x) = \frac{k^2(x\gamma^2 + k - x)}{(x\gamma + k - x)^2}, \quad x \geq 0, \tag{4.3}$$

and

$$q = \begin{cases} \lfloor \frac{k}{1+\gamma} \rfloor, & \text{if } f(\lfloor \frac{k}{1+\gamma} \rfloor) \geq f(\lceil \frac{k}{1+\gamma} \rceil) \text{ and } \lfloor \frac{k}{1+\gamma} \rfloor \geq 1 \\ \lceil \frac{k}{1+\gamma} \rceil, & \text{otherwise.} \end{cases} \tag{4.4}$$

Essentially, q is a maximizer of $f(.)$ over positive integers.

To describe how P_* randomizes the true categories, we represent the true responses with indicator vectors $X = (X_1, \ldots, X_k)$, as in RAPPOR. Then, for a true response $(x_1, \ldots, x_k)$, P_* generates an output $(z_1, \ldots, z_k)$ as follows. Suppose the true category is c_i, which implies $x_i = 1$ and $x_j = 0$ for $j \neq i$. Then, P_* assigns $z_i = 1$ with probability $p = (q\gamma)/(q\gamma + k - q)$ and $z_i = 0$ with probability $1 - p$. Next, if $z_i = 1$, P_* randomly selects $(q-1)$ of the remaining $(k-1)$ components of z and sets those to 1. If $z_i = 0$, P_* assigns 1 to q of the remaining components of z, selected at random. In both cases, all other components of z are 0. Thus, exactly q components of each output vector are 1 and the rest are 0. So, the output space of P_* is $\mathcal{S}_Z^* = \{(z_1, \ldots, z_k) : z_i \text{ is 0 or 1, } i = 1, \ldots, k, \text{ and } \sum z_i = q\}$ and it contains $m = \binom{k}{q}$ elements. As before, let $x^{(i)}$ denote the indicator vector for true response c_i. Then, it can be seen that the transition probabilities of the minimax design are:

$$P((z_1, \ldots, z_k)|x^{(i)}) = \begin{cases} \gamma p_0, & \text{if } z_i = 1 \\ p_0, & \text{if } z_i \neq 1 \end{cases}$$

for $i = 1, \ldots, k$ and $(z_1, \ldots, z_k) \in \mathcal{S}_Z^*$, where $p_0 = k/[\binom{k}{q}(q\gamma + k - q)]$. So, each row of the TPM has two distinct values and parity γ, satisfying the conditions of Theorem 4.1.

As we describe next, the minimax estimator $\tilde{\pi}$ of π under P_* has a simple form and can be calculated easily. With vector representation, both the original and perturbed data under P_* appear as $n \times k$ matrices, with each row showing one respondent's data. Let $\mathbf{V}' = (V_1, \ldots, V_k)$ denote the vector of column sums of the perturbed data matrix. Then, the minimax estimator, derived in Chai and Nayak (2019), is

$$\tilde{\pi} = \frac{(k-1)(q\gamma + k - q)}{q(\gamma - 1)(k - q)}\Big(\frac{\mathbf{V}}{n}\Big) + \frac{1}{k}\Big[\frac{(1-k)(q\gamma + k - q)}{(\gamma - 1)(k - q)} + 1\Big]. \tag{4.5}$$

The minimax estimator is also a method of moments estimator based on $\mathbf{V}$. Here, we want to mention that the estimator in RAPPOR is similar to (4.5). It is based on method of moments and a linear function of the column sums of their perturbed data matrix. Naturally, $\tilde{\pi}$ in (4.5) is an unbiased estimator and it also follows that the risk function, as defined in (4.2), of the mimimax strategy $(P_*, \tilde{\pi})$ is

$$R(P_*, \tilde{\pi}; \pi) = \frac{1}{n}\Big[\frac{(k-1)^2}{f(q) - k} + \frac{1}{k} - 1\Big] + \frac{1}{n}\sum_{i=1}^{k} \pi_i(1 - \pi_i), \tag{4.6}$$

where the function f and the quantity q are as defined in (4.3) and (4.4). Due to unbiasedness (4.6) also gives the trace of the variance-covariance matrix of $\tilde{\pi}$, i.e, $R(P_*, \tilde{\pi}; \pi) = \text{tr}[V(\tilde{\pi})] = \sum V(\tilde{\pi}_i)$. The last term of (4.6) is the risk of the MLE of π under no randomization. So, it reflects only the *sampling variation*. The first term in (4.6) is the *added variance* due to RR, which interestingly is independent of π, unlike the sampling variation.

We want to mention that Duchi et al. (2018) also investigated minimaxity of RR strategies and in a much broader setting. They considered a wider class of problems and loss functions. They derived bounds on minimax values and their convergence rates under ϵ-LDP. In particular, they obtained *rate optimal* methods for certain estimation problems. Naturally, those asymptotic

results may not be useful in small samples. Also, rate optimality ignores the multipliers of convergence rates. So, a rate optimal procedure need not be asymptotically efficient because the minimax risks of two methods may converge to zero at the same (and optimal) rate, but with different multipliers. In contrast, Chai and Nayak's (2019) results are exact, but for a specific problem.

5. Identification Risk Control by Post-randomization

Protecting data confidentiality is a difficult task because disclosure of personal information about survey participants may occur in various forms depending on the context, nature of released data and sensitivity of survey variables. Various types of disclosure and methods for their control are discussed in the books: Willenborg and de Waal (2001), Duncan et al. (2011) and Hundepool et al. (2012). In this article, we shall consider only identity disclosure in microdata release. Consider a complete data set containing values of multiple variables for each of n sampled units. Data agencies commonly publish summaries of the data. But, researchers often want the full data set to explore different models and hypotheses. However, the original data may disclose the values of some sensitive variables for some of the survey participants or units, even if name, social security number and other direct identifiers are removed. In particular, one might be able to correctly identify the records of a target unit by matching gender, race, occupation and other characteristics that can be obtained easily from other sources. Then, one can learn the identified unit's values for all other variables. This is called identity disclosure, which is also regarded as one of the most serious violations of data confidentiality.

5.1. Identification Risk Measures

The variables that an intruder might use for matching are called key (or pseudo-identifying) variables, which are usually categorical. For reducing identification risks, agencies perturb the true values of the key variables and then release the perturbed data. For choosing a suitable

perturbation method, the agency should first determine its disclosure control goals. For that, the agency needs to select and specify the key variables. Thus, we assume that the key variables are given and all are categorical. As before, let X denote the cross-classification of all key variables and suppose X takes values in $\mathcal{S}_X = \{c_1, \ldots, c_k\}$. In this setting, Bethlehem et al. (1990), Skinner and Elliot (2002), Shlomo and De Waal (2008), Shlomo and Skinner (2010) and others have proposed and investigated different measures of identification risk.

Early works focused mainly on the units that are unique in the sample with respect to X. As agencies do not reveal which population units are in the sample, it is reasonable to assume that an intruder would not know if his target is in the sample or not. Such an intruder would correctly match a sample unique unit if it is also population unique with respect to X. Motivated by this, Bethlehem et al. (1990) defined identification risk as the probability that a unit is population unique, given that it is sample unique, both with respect to X. Obviously, this concerns only the sample unique units. Also, it ignores the effects of data perturbation. So, this measure is not useful for determining a suitable perturbation mechanism. Essentially, it aims to assess how much protection the sample unique units get from sampling. We refer to Skinner and Elliot (2002) for a discussion of similar measures and related references.

Shlomo and Skinner (2010) took a more relevant approach that focuses on correct matches in perturbed (and released) data. Naturally, they take data perturbation into account. However, they were concerned only with the unique matches in released data, presuming those to be the worst cases. Consequently, they defined a unit's identification risk as the probability that the unit is correctly identified given that it has a unique match in released data. This (conditional) probability is with respect to both sampling and data perturbation. This identification risk is unit specific and it varies over the sampled units.

There are some practical difficulties in using the preceding risk measure in finding a suitable perturbation mechanism. First, the risks of the sampled units involve the unknown population frequencies and hence those cannot be calculated from available information. Methods for estimating those have been proposed, but they require assumptions about the population dis-

tribution and data modeling. So, the estimates depend on the model assumptions. Second, the effectiveness of data perturbation is assessed using the average of the risks of all sampled units. That may not be appropriate because a small average risk does not imply that disclosure risks of all units are desirably small. Third, the search for a suitable perturbation procedure requires an iterative approach. One would need to assess the effectiveness of several procedures to select a procedure. For example, to apply data swapping, as described in Shlomo and Skinner (2010), one would need to evaluate the average risk measure for various swap rates to choose a suitable value for actual application.

Recently, Nayak, Zhang and You (2018), henceforth NZY, refined Shlomo and Skinner's (2010) approach and introduced a strict identification risk control goal. They also developed a method for achieving that goal. To describe the NZY approach, consider an intruder J who wants to identify the records of a target unit B in the released perturbed data. Let $X_{(B)}$ denote B's value of X, and suppose $X_{(B)} = c_j$. NZY assumed that (a) J knows $X_{(B)}$, (b) J knows that B is in the sample and (c) J randomly selects one of the records in the released data that match $X_{(B)}$, and identifies those as B's data. If no records in released data match $X_{(B)}$, the intruder stops his search for B's data. While assumptions (a) and (c) are realistic, (b) is overly stringent because agencies do not disclose which population units are included in the sample.

Let T_j and S_j denote the frequencies of c_j in the original and perturbed data, respectively, and let $\mathbf{T} = (T_1, ..., T_k)'$ and $\mathbf{S} = (S_1, ..., S_k)'$. Note that if $X_{(B)} = c_j$, then S_j records in the released data match B on key variables. Intuitively, J's confidence in a declared match depends on S_j. Observing this, NZY considered the following to propose a strict disclosure control goal:

$$R_j(a) = P(CM|X_{(B)} = c_j, S_j = a), \quad j = 1, \ldots, k, a \geq 1$$

where CM denotes the event that B is *correctly matched* in the preceding setup. NZY proposed that the agency should select a suitable value ξ and guarantee, with appropriate data

perturbation, that

$$R_j(a) \leq \xi \quad \text{for all } j = 1, \ldots, k, \text{ and all integers } a \geq 1. \tag{5.1}$$

Then, no unit's correct match probability would exceed ξ. This gives a clear and strong identification risk control goal.

Like all past identification risk measures, $R_j(a)$ also depends on the unknown population frequencies. So, we cannot calculate $R_j(a)$'s and thereby verify whether a data perturbation mechanism guarantees (5.1) or not. To avoid this difficulty, NZY considered

$$R_j(a, \mathbf{t}) = P(CM | X_{(B)} = c_j, S_j = a, \mathbf{T} = \mathbf{t}),$$

further conditioning on $\mathbf{t}$. Quite importantly, $R_j(a, \mathbf{t})$'s do not involve unknown parameters under PRAM and so, those can be assessed and controlled without estimating any parameter. Note that $R_j(a, \mathbf{t})$'s involve the original frequency vector $\mathbf{t}$, but that is available to the data agency. NZY suggested to satisfy (5.1) by using a data perturbation mechanism such that

$$R_j(a, \mathbf{t}) \leq \xi \quad \text{for } j = 1, \ldots, k, \text{ all } a > 0 \text{ and all } \mathbf{t}. \tag{5.2}$$

Effectively, (5.2) is their disclosure control goal, which readily implies (5.1).

5.2. A Post-randomization Method

Taking the preceding approach, NZY developed a class of unbiased PRAMs that can be used to satisfy (5.2) for $\xi > 1/3$. They give two reasons for choosing a $\xi > 1/3$ in practical situations. First, intruders should have strong evidence for declaring matches. To be credible, the correct match probability for a declared match should be substantial, perhaps larger than 0.5. Second, as noted earlier, assumption (b) is overly stringent. Usually, an intruder would not know if a target is in the sample or not. For such an intruder, a correct match probability is much

smaller than $R_j(a)$, approximately $R_j(a)$ times the target's sample inclusion probability, which is usually quite small.

For any given $\xi > 1/3$, NZY developed an unbiased PRAM to satisfy (5.2) as follows. First, we should mention that any unbiased PRAM does not affect the empty categories. In other words, an unbiased PRAM does not change a true category to a category that was originally empty. Truly, $\mathcal{S}_Z$ consists of only the categories in $\mathcal{S}_X$ that have positive frequencies in the original data set, i.e., $\mathcal{S}_Z = \mathcal{S}_X^* = \{c_i : c_i \in \mathcal{S}_X \text{ and } t_i > 0\}$, which may be a proper subset of $\mathcal{S}_X$. Actually, $\mathcal{S}_X^*$ is also the input space as all observed values are in this set. For notational simplicity, we assume that all categories are nonempty and thus $\mathcal{S}_X^* = \mathcal{S}_X$.

The NZY method uses one specific class of unbiased PRAMs. Specifically, they use the transition probabilities

$$p_{ij} = P(Z = c_i | X = c_j) = \begin{cases} 1 - \frac{\theta}{t_j}, & \text{if } i = j; \\ \frac{\theta}{(k-1)t_j}, & \text{if } i \neq j, \end{cases} \tag{5.3}$$

where t_j is the original frequency of c_j and θ is a design parameter, chosen suitably to satisfy (5.2). Clearly, the TPM $P = ((p_{ij}))$ given by (5.3) is adaptive, viz. it depends on the observed data via the category frequencies. Also, P has a simple structure. It changes a true category c_j with probability θ/t_j, which is inversely proportional to the frequency of the unit's true category. If a true category is c_j, then it is kept unchanged with probability $1-\theta/t_j$. When a true category is changed, the replacement is selected at random from the remaining categories. One helpful feature of this P is that it is determined fully by a single parameter θ. So the effects of P on identification risks and statistical inferences can be studied in terms of θ only.

One key result of NZY is that for given $1/3 < \xi < 1$, the above P satisfies (5.2) if θ is chosen

as the solution of $h(\theta) = \xi$, where

$$h(\theta) = \begin{cases} \frac{1-\theta}{1-\theta+\theta^2}, & \text{if } \theta \leq \frac{2}{3}, \\ \frac{2-\theta}{4-2\theta+\theta^2}, & \text{if } \theta > \frac{2}{3}, \end{cases}$$

and $k \geq (1-\theta)^{-1}$. They also showed that $h(\theta)$ is a strictly decreasing function of θ, with $h(0) = 1$ and $h(1) = 1/3$ and thus for any $1/3 < \xi < 1$, $h(\theta) = \xi$ admits a unique solution for θ in $(0, 1)$. This result gives a theoretical basis for designing a post-randomization method for guaranteeing (5.2). Also, a suitable PRAM can be designed directly, without iterative calculations or adjustments, unlike previous approaches.

Actually, the method proposed by NZY applies the preceding result separately to subsets of the data set, which are formed by partitioning the data into homogeneous groups and then taking only the sensitive records in each group. That is done to better preserve data utility. While category and cell are synonymous, in the rest of this section we shall use cell for a cross-classified variable and category for individual variables, for additional clarity. So, we shall use cell for X, as it is the cross classification of all key variables. For given $0 < \xi < 1$, a cell is considered *sensitive* if its frequency is less than $1/\xi$. A cell c_j is nonsensitive if $t_j \geq 1/\xi$ because in the original data, the probability of correctly identifying a unit falling in that cell is $1/t_j \leq \xi$. The identification risks of all units in the nonsensitive cells are already sufficiently small. So, we only need to post-randomize X for all units in the sensitive cells. All sensitive cells in a partition set form a post-randomization block (PRB). The NZY method applies PRAM to the PRB's separately. More details of the method and some parts of an illustrative example are given below.

The main purpose of data partitioning is to control the nature and magnitude of possible changes due to PRAM, and even preserve selected parts and summaries of the data set. NZY gave several ideas for data partitioning. One simple approach is to partition the data by broader or generalized categories of the key variables. As an illustrative example, NZY applied the

method to a data set publicly released by the U.S. Census Bureau. It contains values of several demographic and economic variables for 59,033 individuals. For illustration, NZY took gender (2), age (92), race (9), marital status (5) and Public Use Microdata Area (PUMA) (44) as the key variables, where the values in parentheses show the number of categories of the variables. The cross-classification of these key variables yields 364,320 cells.

In the example, NZY partitioned the data by gender, seven age intervals, viz. 0–17, 18–24, 25–34, 35–44, 45–54, 55–64, and 65 and above, and the three race categories: white, black and 'other races.' That divided the data into 42 partition sets, corresponding to all possible combinations of gender, 7 age intervals and 3 race classes. For example, all females of 'other races' with age between 25 and 34 constitute one partition set. Similarly, all white males in the age interval 55–64 form another partition set. Note for example that all individuals in a partition set are either male or female. As the method applies PRAM within each partition set, it will not alter the gender of any individual. Similarly, it will preserve race if the original category is white or black, which are the two major categories. Race will change only among the other races. Age will remain in the partitioning intervals. For example, if the true value is 38, the perturbed value will be between 35 and 44. It will preserve the counts of voting age (18 or above) and senior (65 and above) people, which are important in policy research. Also note that since marital status and PUMA were not used in data partitioning, those may change freely. This partition in one extreme fully preserves gender and on the other extreme permits unlimited changes of marital status and PUMA.

In the example, NZY took $\xi = 0.395$, for which only singleton and doubleton cells (with frequency 1 and 2, respectively) are sensitive. So, all singleton and doubleton cells of X in a partition set formed one PRB. In each of the 42 PRB's, the true X cells are post-randomized using the transition probabilities given by (5.3). For $\xi = 0.395$, it turns out that $\theta = 0.8$ and $(1-\theta)^{-1} = 5$. Earlier, we discussed (5.3) for one data set and assuming that all cells are nonempty. For applying post-randomization, we need to specialize (5.3) for each PRB. Specifically, we need to interpret k as the number of cells in the PRB, which changes from PRB

to PRB. Also, $c_1, \ldots, c_k$ should represent the cells within a PRB. The theoretical results for guaranteeing (5.2) also require at least $(1-\theta)^{-1}(= 5$ for $\xi = 0.395)$ cells in each PRB. This was satisfied in the example. Actually, the number of cells in the 42 PRB's ranged between 124 and 1480. One should not partition the data overly finely into too many sets so that the condition $k \geq (1-\theta)^{-1}$ is satisfied.

We mention some other facts from the NZY example. The five key variables defined 364,320 cells. The data set, with sample size 59,033, showed only 25,406 nonempty cells, of which 13,662 are singleton and 4,777 are doubleton. As $\theta = 0.8$, the method changed the true cell of each singleton unit with probability 0.8 and each doubleton unit with probability 0.4. When a true cell was changed, the new cell was picked at random from the remaining cells within the PRB. The method kept the true values of all nonsensitive units unchanged.

Deriving methods for analyzing perturbed data, making appropriate adjustments for data perturbation, is burdensome to data users. Also, data users usually do not get full information about the perturbation mechanism that is needed for modeling the perturbation effects. So, it is important to perturb the data in such a way that standard inferential methods for the original data remain valid for the released data, at least approximately. Generally, we want perturbation methods that add a small (or negligible) variance and no bias. The NZY methods does quite well in that respect, largely due to data partitioning and using unbiased PRAM. The relative frequencies based on perturbed data are unbiased estimators of corresponding population probabilities. NZY examined the variance of these estimators and proved that the additional variance due to data perturbation is of order $1/n^2$, where n is sample size. That is negligible in comparison to sampling variance, which is of order $1/n$.

Consistent with the theoretical results, the NZY method exhibited very small effects on data distributions in their example. We reproduce the distributions of marital status and race based on the original and perturbed data in Tables 1 and 2. There, columns 2 and 3 give the original and perturbed frequencies and the numbers in parentheses are relative frequencies. The last column gives the difference between the original and perturbed frequencies. Recall that marital

status was allowed to change freely. Even then, the original and perturbed frequencies are very close. The differences between original and perturbed frequencies of race categories are also quite small. Note that the difference is 0 for white and black. That is not by coincidence, but due to the particular data partitioning, which forced to preserve race for those two groups. We refer interested readers to the NZY paper for more details about the method and the example.

Table 1: Frequency Distributions of Marital Status

Marital Status	Original Data	Perturbed Data	Difference
Married	24688 (.4182)	24678 (.4180)	10
Widowed	3156 (.0535)	3180 (.0539)	-24
Divorced	4742 (.0803)	4704 (.0797)	38
Seperated	1040 (.0176)	1039 (.0176)	1
Never married	25407 (.4304)	25432 (.4308)	-25

Table 2: Frequency Distribution of Race

Race	Original Data		Perturbed Data		Difference
White	37201	(.6302)	37201	(.6302)	0
Black	15239	(.2581)	15239	(.2581)	0
American Indian alone	97	(.0016)	92	(.0015)	5
Alaska Native alone	1	(.00002)	0	(0)	1
American Indian & Alaska Native	42	(.0007)	46	(.0008)	-4
Asian	3461	(.0586)	3445	(.0584)	16
Native Hawaiian & other Pacific Islander	20	(.0004)	21	(.0004)	-1
Some other race alone	1349	(.0228)	1337	(.0227)	12
Two or more races	1623	(.0275)	1652	(.0280)	-29

6. Discussion

The idea of randomizing true responses for protecting respondent's privacy and data confidentiality has been around for a long time. But, it has not been used much in real surveys, perhaps due to lack of practical privacy measures and adequate guidance on choosing the transition probabilities. In recent years, RR methods have received significant attention from companies and computer scientists in a new context, viz. for protecting privacy when recording data from

various online activities. Recent research has yielded precise privacy concepts and measures and rigorous methods for determining the transition probabilities. We have reviewed some of those developments. In particular, we covered one approach to strict privacy protection (in Sections 3 and 4) and one rigorous method for identification risk control in releasing microdata (in Section 5).

RR surveys and PRAM are similar in that both randomize true responses with predetermined probabilities and the transition probabilities govern their mathematical properties. But, one important difference is that in PRAM, the original data (containing true responses of all units) may be used to choose the transition probabilities, whereas in RR surveys, those must be determined before data collection. This implies that in PRAM, randomization may be applied after data partitioning and only to some selected units. The NZY method displays and utilizes these special features of PRAM.

Privacy and data confidentially are difficult but important topics and have been investigated for a long time. Also, research in these areas has increased significantly in recent years. New theories and methods are being developed by researchers in statistics, computer science, public policy and other fields. Other concepts and methods such as grouping, data swapping, synthetic data, l-diversity and differential privacy have been developed to mitigate disclosure risks. We consider response randomization as one of the most basic and promising tools for protecting privacy and data confidentiality. In particular, we believe that there is substantial scope for developing post-randomization methods, like the NZY method, for protecting data confidentiality.

Acknowledgment. The author thanks Eric Slud, Tommy Wright, Bimal Sinha and Kyle Erimata for reading an earlier draft and giving many suggestions, which have helped to improve the paper significantly.

References

[1] Agrawal, S., Haritsa, J.R. and Prakash, B.A. (2009). FRAPP: a framework for high-accuracy privacy-preserving mining. *Data Mining and Knowledge Discovery*, 18, 101-139.

[2] Aggarwal, C.C. and Yu, P.S. (Eds.) (2008). *Privacy-Preserving Data Mining: Models and Algorithms*, New York: Springer Science and Business Media.

[3] Bethlehem, J.G., Keller, W.J., and Pannekoek, J. (1990). Disclosure control of microdata, *Journal of the American Statistical Association*, 85, 38-45.

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

[5] Blackwell, D. (1953). Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24, 265-272.

[6] Blair, G., Imai, K. and Zhou, Y-Y. (2015). Design and analysis of the randomized response technique, *Journal of the American Statistical Association*, 110, 1304-1319.

[7] Chai, J., and Nayak, T.K. (2018). A criterion for privacy protection in data collection and its attainment via randomized response procedures. *Electronic Journal of Statistics*, 12, 4264-4287.

[8] Chai, J., and Nayak, T.K. (2019). Minimax randomized response methods for providing local differential privacy. Research Report Series, Statistics #2019-04, Center for Statistical Research & Methodology, U.S. Census Bureau.
https://www.census.gov/srd/papers/pdf/RRS2019-04.pdf

[9] Chaudhuri, A. (2001). Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population. *Journal of Statistical Planning and Inference*, 94, 37-42.

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

[11] Chaudhuri, A. (2010). *Randomized Response and Indirect Questioning Techniques in Surveys*. Boca Raton: CRC Press.

[12] Chaudhuri, A., Christofides, T.C. and Rao, C.R. (editors). (2016). *Data Gathering, Analysis and Protection of Privacy through Randomized Response Techniques: Qualitative and Quantitative Human Traits*, New York: Elsevier .

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

[14] Chen, B-C., Kifer, D., LeFevre, K. and Machanavajjhala, A. (2009). Privacy-preserving data publishing. *Foundations and Trends in Databases*, 2, 1-167.

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

[16] Cormode, G., Jha, S., Kulkarni, T., Li, N., Srivastava, S. and Wang, T. (2018) . Privacy at scale: Local differential privacy in practice. In *Proceedings of the 2018 International Conference on Management of Data*, ACM, 1655-1658.

[17] Ding, B., Kulkarni, J. and Yekhanin, S. (2017). Collecting telemetry data privately. In *Advances in Neural Information Processing Systems 30*, pp. 3574-3583.

[18] Duchi, J.C., Jordan, M.I., and Wainwright, M.J. (2018). Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113, 182-201.

[19] Duncan, G.T., Elliot, E. and Juan Jose Salazar, G. (2011). *Statistical Confidentiality: Principles and Practice*, New York: Springer.

[20] Evfimievski, A., Gehrke, J. and Srikant, R. (2003). Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the Twenty-Second Symposium on Principles of Database Systems*, ACM, pp. 211-222.

[21] Erlingsson, U., Pihur, V., and Korolova, A. (2014). Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, ACM, 1054-1067.

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

[23] Fung, B.C.M., Wang, K., Fu, A.W-C. and Yu, P.S. (2019). *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. New York: CRC Press.

[24] Gouweleeuw, J.M., Kooiman, P., and de Wolf, P.P. (1998). Post randomisation for statistical disclosure control: Theory and implementation. *Journal of Official Statistics*, 14, 463-478.

[25] Holbrook, A.L. and Krosnick, J.A. (2010). Measuring voter turnout by using the randomized response technique: Evidence calling into question the method's validity. *Public Opinion Quarterly*, 74, 328-343.

[26] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K. and de Wolf, P-P. (2012). *Statistical Disclosure Control*, New York: Wiley.

[27] Kairouz, P., Oh, S., and Viswanath, P. (2016). Extremal mechanisms for local differential privacy. *Journal of Machine Learning Research*, 17, 1-51.

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

[29] Lanke, J. (1976). On the degree of protection in randomized interviews. *International Statistical Review*, 44, 197-203.

[30] Leysieffer, R.W. and Warner, S.L. (1976). Respondent jeopardy and optimal designs in randomized response models. *Journal of the American Statistical Association*, 71, 649-656.

[31] Nayak, T.K. (1994). On randomized response surveys for estimating a proportion. *Communications in Statistics - Theory and Methods*, 23, 3303-3321.

[32] Nayak, T.K. and Adeshiyan, S.A. (2009). A unified framework for analysis and comparison of randomized response surveys of binary characteristics. *Journal of Statistical Planning and Inference*, 139, 2757-2766.

[33] Nayak, T.K. and Adeshiyan, S.A. (2016). On invariant post-randomization for statistical disclosure control, *International Statistical Review*, 84, 26-42.

[34] Nayak, T.K., Adeshiyan, S.A. and Zhang, C. (2016). A concise theory of randomized response techniques for privacy and confidentiality protection, in A. Chaudhuri, T.C. Christofides and C.R. Rao editors, *Data Gathering, Analysis and Protection of Privacy through Randomized Response Techniques: Qualitative and Quantitative Human Traits*, New York: Elsevier, pp. 273-286.

[35] Nayak, T.K., Zhang, C., and Adeshiyan, S.A. (2015). Emerging applications of randomized response concepts and some related issues. *Model Assisted Statistics and Applications*, 10, 335-344.

[36] Nayak, T.K., Zhang, C., and You, J. (2018). Measuring identification risk in microdata release and its control by post-randomisation. *International Statistical Review*, 86, 300-321.

[37] Padmawar, V.R. and Vijayan, K. (2000). Randomized response revisited. *Journal of Statistical Planning and Inference*, 90, 293-304.

[38] Shlomo, N. and Skinner, C., 2010. Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata. *The Annals of Applied Statistics*, 4, 1291-1310.

[39] Shlomo, N., and De Waal, T. (2008). Protection of micro-data subject to edit constraints against statistical disclosure. *Journal of Official Statistics*, 24, 229-253.

[40] Skinner, C.J., and Elliot, M. J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society, Ser. B*, 64, 855-867.

[41] Van den Hout, A. and Elamir, E.A.H., 2006. Statistical disclosure control using post randomisation: variants and measures for disclosure risk. *Journal of Official Statistics*, 22, 711-731.

[42] Van den Hout, A. and Van der Heijden, P.G.M., 2002. Randomized response, statistical disclosure control and misclassification: a review. *International Statistical Review*, 70, 269-288.

[43] Wang, S., Huang, L., Wang, P., Nie, Y., Xu, H., Yang, W., Li, X-Y. and Qiao, C. (2016). Mutual information optimally local private discrete distribution estimation. *arXiv preprint arXiv:1607.08025*

[44] Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.

[45] Willenborg, L.C.R.J. and De Waal, T. (2001). *Elements of Statistical Disclosure Control*, New York: Springer.

[46] Ye, M. and Barg, A. (2018). Optimal schemes for discrete distribution estimation under locally differential privacy. *IEEE Transactions on Information Theory*, 64, 5662-5676.

[47] Zhimin, H. and Zaizai, Y. (2012). Measure of privacy in randomized response model. *Quality & Quantity*, 46, 1167-1180.