Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

A standardization of randomized response strategies

by Andreas Quatember

December 2009

Statistics Canada Statistique Canada

Canada

Survey Methodology, December 2009
Vol. 35, No. 2, pp. 143-152
Statistics Canada, Catalogue No. 12-001-X

A standardization of randomized response strategies

Andreas Quatember [1]

Abstract

Randomized response strategies, which have originally been developed as statistical methods to reduce nonresponse as well as untruthful answering, can also be applied in the field of statistical disclosure control for public use microdata files. In this paper a standardization of randomized response techniques for the estimation of proportions of identifying or sensitive attributes is presented. The statistical properties of the standardized estimator are derived for general probability sampling. In order to analyse the effect of different choices of the method's implicit "design parameters" on the performance of the estimator we have to include measures of privacy protection in our considerations. These yield variance-optimum design parameters given a certain level of privacy protection. To this end the variables have to be classified into different categories of sensitivity. A real-data example applies the technique in a survey on academic cheating behaviour.

Key Words: Privacy protection; Statistical disclosure control; Nonresponse; Untruthful answering.

1. Introduction

The occurence of nonresponse and the unwillingness to provide the true answers are natural in survey sampling. They may result in an estimator of population parameters, which has a bias of unknown magnitude and a high variance. A responsible user therefore cannot ignore the presence of nonresponse and untruthful answering.

Let U be the universe of N population units and U_A be a subset of N_A elements, that belong to a class A of a categorial variable under study. Moreover let U_A^c be the group of N_A^c elements, that do not belong to this class $(U = U_A \cup U_A^c, U_A \cap U_A^c = \emptyset, N = N_A + N_A^c)$. Let

$$x_i = \begin{cases} 1 & \text{if unit } i \in U_A, \\ 0 & \text{otherwise} \end{cases}$$

$(i = 1, 2, ..., N)$ and the parameter of interest be the relative size π_A of subpopulation U_A:

$$\pi_A = \frac{\sum_U x_i}{N} = \frac{N_A}{N} \qquad (1)$$

$(\sum_U x_i$ is abbreviated notation for $\sum_{i \in U} x_i)$. In a probability sample s (see for instance: Särndal, Swensson and Wretman 1992, page 8f) an estimator of π_A can be calculated from the Horvitz-Thompson estimator of N_A by

$$\hat{\pi}_A^{\text{dir}} = \frac{1}{N} \cdot \sum_s \frac{x_i}{\pi_i} \qquad (2)$$

$(\pi_i > 0$ is the probability that unit i will be included in the sample), if the question "Are you a member of group U_A?" (or an equivalent question) is asked directly (dir). This estimator is unbiased, if all x_i's $(i = 1, 2, ..., n)$ are observed truthfully. In the presence of unit or item nonresponse with respect to a variable under study the sample s is divided into a "response set" $r \subset s$ of size n_r and a "missing set" $m \subset s$ of size n_m $(s = r \cup m, r \cap m = \emptyset, n = n_r + n_m)$. For variables of a highly personal, embarrassing matter (like drug addiction, diseases, sexual behaviour, tax evasion, alcoholism, domestic violence or involvement in crimes) r is furthermore divided into a set t of n_t sample units, who answer truthfully, and a set u of size n_u, who answer untruthfully $(r = t \cup u, t \cap u = \emptyset, n_r = n_t + n_u)$. Estimator (2) must then be rewritten as:

$$\hat{\pi}_A^{\text{dir}} = \frac{1}{N} \cdot \left(\sum_t \frac{x_i}{\pi_i} + \sum_u \frac{x_i}{\pi_i} + \sum_m \frac{x_i}{\pi_i} \right). \qquad (3)$$

Evidently the elements of set u cannot be identified and the x_i's of m are not observable. This imposes errors of measurement and nonreponse on the estimation. Therefore everything should be done to keep the untruthful answering rate as well as the nonresponse rate as low as possible.

Survey design features, which clearly affect both the quantity and the quality of the information asked from the respondents (see for instance: Groves, Fowler, Couper, Lepkowski, Singer and Tourangeau 2004, Section 6.7), are strongly related to the sample units' concerns about "data confidentiality" and "perceived protection of privacy". The first term refers to the respondents' desire to keep replies out of hands of uninvolved persons, whereas the second refers to the wish to withhold information from absolutely anybody. Singer, Mathiowetz and Couper (1993) and Singer, van Hoewyk and Neugebauer (2003) report on two successive U.S. population surveys, that the higher these concerns are the lower is the probability of the respondent's participation in the survey (page 470ff and page 375ff).

1. Andreas Quatember is Assistant Professor at the IFAS-Department of Applied Statistics, Johannes Kepler University Linz, Altenberger Str. 69, A-4040 Linz, Austria, Europe. Web address: www.ifas.jku.at. E-mail: andreas.quatember@jku.at.

What can statisticians contribute to this important field of research? For awkward questions the use of *randomized response strategies* at the survey's design stage may reduce the rates of nonresponse and of untruthful answering due to a perceived increase of privacy protection. A common characteristic of these methods is that instead of the direct questioning on the sensitive subject a questioning design is used, which does not enable the data collector to identify the (randomly selected) question on which the respondent has given the answer, although it does still allow to estimate the parameter under study. The idea is to reduce in this way the individuals' fear of an embarrassing "outing" to make sure that the responding person is willing to cooperate. To achieve this goal the respondent clearly has to understand how the questioning design does protect his or her privacy (*cf.* Landsheer, van der Heijden and van Gils 1999, page 6ff).

Pioneering work in this field was published by Warner (1965). In his questioning design each respondent has to answer randomly either with probability p_1 the question "Are you a member of group U_A?" or with probability $p_2 = 1 - p_1$ the alternative "Are you a member of group U_A^c?" $(0 < p_1 < 1)$. Since then various randomized response techniques with differing randomization devices have been proposed (for a review see: Chaudhuri and Mukerjee 1987, Nathan 1988 or Tracy and Mangat 1996). All of these strategies make use of randomly selected questions or answers, though some of them use different random devices depending on the respondent's possession or nonpossession of a certain attribute (see for example: Kuk 1990; Mangat 1994; Kim and Warde 2005).

Warner (1971) was the first to note that these techniques are also applicable as methods of masking confidential micro-data sets to allow their release for public use (*cf.* ibd., page 887). Such microdata sets might contain variables, which allow the direct identification of survey units like the name or an identification number, but also variables, which contain sensitive information on an individual. To protect the survey units against disclosure it might not suffice to delete the variables, which are directly linked to entities, because some of the units might still be identifiable by the rest of their records. Statistical disclosure control is nothing else but a balancing act between the protection of the anonymity of the survey units and the preservation of information contained in the data (*cf.* Skinner, Marsh, Openshaw and Wymer 1994). Methods of data masking can be classified into three categories (*cf.* Domingo-Ferrer and Mateo-Sanz 2002 or Winkler 2004): (1) The *global recoding* of variables into less detailed categories or larger intervals (see for instance: Willenborg and de Waal 1996, page 5f) or the *local recoding* using different grouping schemes at unit level (*cf.* Hua and Pei 2008, page 215f). (2) The *local suppression* of certain variables for survey units with a high risk of re-identification by simply setting their values at "missing" (*cf.* Willenborg and de Waal 1996, page 77). (3) The *substitution* of true values of a variable by other values.

One of the strategies of the third category is the *micro-aggregation* of variables (*cf.* Defays and Anwar 1998). Therein the true variable values are for example sorted by size and then divided into (small) groups. Within each group data aggregates are released instead of the original observations. Another such method is *data-swapping*, where data from units with a high risk of re-identification are interchanged with data from another subset of survey units (*cf.* Dalenius and Reiss 1982). Another technique of substituting identifying or sensitive information is the *addition of noise* to the observed values, meaning that the outcome of a random experiment is added to each datum (*cf.* Dalenius 1977 or Fuller 1993). Finally also the randomized response techniques can be used to mask identifying or sensitive variables. In this case either the survey units already perform the data masking at the survey's design stage or the statistical agency applies the probability mechanism of the technique before the release of the microdata file (*cf.* Rosenberg 1980, Kim 1987, Gouweleeuw, Kooiman, Willenborg and de Wolf 1998, or van den Hout and van der Heijden 2002).

All methods of statistical disclosure control protect the survey units' privacy by a loss of information, which can be seen as the price that has to be paid for it. To be able to appropriately adjust the estimation process the user of the microdata file has to be informed about the details of the masking procedure.

A new standardization of the techniques of randomized response follows in Section 2 of this paper. Furthermore the statistical properties of the standardized estimator are derived for general probability sampling. In Section 3 the essential perspective of privacy protection is described. The question, which of the special cases included in the standardization is most efficient, is answered in the subsequent Section 4. Section 5 contains a real-data example, which demonstrates the application of the recommendations of Section 4 in a survey on academic cheating behaviour.

2. Standardizing randomized response strategies

Let us formulate the following standardization of the randomized response strategies: Each respondent has either to answer randomly with probability

- p_1 the question "Are you a member of group U_A?",
- p_2 the question "Are you a member of group U_A^c?"

or

– p_3 the question "Are you a member of group U_B? "

or is instructed just to say

– "yes" with probability p_4 or

– "no" with probability p_5

($\sum_{i=1}^{5} p_i = 1,\ 0 \le p_i \le 1$ for $i = 1, 2, ..., 5$). The N_B elements of group U_B are characterized by the possession of a completely innocuous attribute B (for instance a season B of birth), that should not be related to the possession or nonpossession of attribute A. This nonsensitive question on membership of group U_B was introduced as an alternative to the question on membership of U_A by Horvitz, Shah and Simmons (1967) to further reduce the respondent's perception of the sensitivity of the procedure. $\pi_B = N_B/N$ (with $0 < \pi_B < 1$) is the relative size of group U_B .π_B and the probabilities $p_1, p_2, ..., p_5$ are the *design parameters* of our standardized randomized response technique.

Let

$$y_i = \begin{cases} 1 & \text{if unit } i \text{ answers "yes",} \\ 0 & \text{otherwise} \end{cases}$$

($i = 1, 2, ..., n$). For an element i the probability of a "yes"-answer with respect to the randomized response questioning design R is for given x:

$$P_R(y_i = 1) = p_1 \cdot x_i + p_2 \cdot (1 - x_i) + p_3 \cdot \pi_B + p_4 = a \cdot x_i + b \qquad (4)$$

with $a \equiv p_1 - p_2$ and $b \equiv p_2 + p_3 \cdot \pi_B + p_4$. Then the term

$$\hat{x}_i = \frac{y_i - b}{a}$$

is unbiased for the true value $x_i (a \neq 0)$. Using these "substitutes" for x_i (and assuming full cooperation of the respondents) the following theorems apply:

Theorem 1: For a probability sampling design with inclusion probabilities π_i the following unbiased estimator of parameter π_A is given:

$$\hat{\pi}_A = \frac{1}{N} \cdot \sum_s \frac{\hat{x}_i}{\pi_i}. \qquad (5)$$

Theorem 2: For a probability sampling design P the variance of the standardized estimator $\hat{\pi}_A$ (5) is given by

$$V_P(\hat{\pi}_A) = \frac{1}{N^2} \cdot \left(V_P\left(\sum_s \frac{x_i}{\pi_i} \right) + \frac{b \cdot (1-b)}{a^2} \cdot \sum_U \frac{1}{\pi_i} + \frac{1 - 2 \cdot b - a}{a} \cdot \sum_U \frac{x_i}{\pi_i} \right). \qquad (6)$$

For the proofs of both theorems see the Appendix. The first summand within the outer brackets of (6) refers to the variance of the Horvitz-Thompson estimator for the total $\sum_U x_i$ for a probability sampling design P when the question on membership of U_A is asked directly. The second one can be seen as the price we have to pay in terms of accuracy for the privacy protection offered by the randomized response questioning design. Apparently this variance can be estimated unbiasedly by inserting an unbiased estimator $\hat{V}_P(\sum_s x_i / \pi_i)$ for $V_P(\sum_s x_i / \pi_i)$ and $\sum_s \hat{x}_i / \pi_i^2$ for $\sum_U x_i / \pi_i$.

For simple random sampling without replacement for instance estimator (5) is given by

$$\hat{\pi}_A = \frac{\hat{\pi}_y - b}{a} \qquad (7)$$

with $\hat{\pi}_y = \sum_s y_i / n$, the proportion of "yes"-answers in the sample. In this case the variance (6) of the standardized estimator $\hat{\pi}_A$ is given by

$$V(\hat{\pi}_A) = \frac{\pi_A \cdot (1 - \pi_A)}{n} \cdot \frac{N - n}{N - 1} + \frac{1}{n} \cdot \left(\frac{b \cdot (1-b)}{a^2} + \frac{1 - 2 \cdot b - a}{a} \cdot \pi_A \right). \qquad (8)$$

This theoretical variance is unbiasedly estimated by

$$\hat{V}(\hat{\pi}_A) = \frac{\hat{\pi}_A \cdot (1 - \hat{\pi}_A)}{n - 1} \cdot \frac{N - n}{N} + \frac{1}{n} \cdot \left(\frac{b \cdot (1-b)}{a^2} + \frac{1 - 2 \cdot b - a}{a} \cdot \hat{\pi}_A \right). \qquad (9)$$

To be able to calculate $\hat{\pi}_A$ at all, the question on membership of U_A (or U_A^c, but we will ignore this possibility subsequently without loss of generality) must be included in the questioning design with $p_1 > 0$. There is a total of 16 combinations of this question with the four other questions or answers (see: Table 1). These combinations can be described as special cases of our standardized response strategy. For example choosing $p_1 = 1$ leads to the direct questioning on the subject. If we let $0 < p_1 < 1$ and $p_2 = 1 - p_1$ the standardized questioning design turns into Warner's procedure. For $0 < p_1 < 1$ and $p_3 = 1 - p_1$ one gets Horvitz *et al.*'s technique with known π_B (see: Greenberg, Abul-Ela, Simmons and Horvitz 1969). (For other special cases already published as to the best of our knowledge, the reader is referred to the "References"-column of Table 1).

The question, that arises directly from these considerations, is how to choose the design parameters of the standardized response technique to find out the strategies that perform best. We will answer this question in Section 4. But for this purpose we have to include the level of privacy protection, which results from choosing these parameters differently, in our considerations.

Table 1
All special cases of the standardized randomized response strategy

Design	Questions/Answers					References
	U_A	U_{A^c}	U_B	yes	no	
*ST*1	•					Direct questioning
*ST*2	•	•				Warner (1965)[1]
*ST*3	•		•			Greenberg *et al.* (1969)[2]
*ST*4	•			•		
*ST*5	•				•	
*ST*6	•	•	•			
*ST*7	•	•		•		
*ST*8	•	•			•	Quatember (2007)[3]
*ST*9	•		•	•		
*ST*10	•		•		•	Singh, Horn, Singh and Mangat (2003)[4]
*ST*11	•			•	•	Fidler and Kleinknecht (1977)[5]
*ST*12	•	•	•	•		
*ST*13	•	•	•		•	
*ST*14	•	•		•	•	
*ST*15	•		•	•	•	
*ST*16	•	•	•	•	•	

1. A two-stage version was presented by Mangat and Singh (1990)
2. A two-stage version was presented by Mangat (1992)
3. This is a one-stage version of Mangat, Singh and Singh (1993)
4. This is a one-stage version of Singh, Singh, Mangat and Tracy (1994)
5. A two-stage version was presented by Singh, Singh, Mangat and Tracy (1995)

3. Privacy protection

To be able to compare the efficiency of questioning designs with different design parameters it is apparently inevitable to measure the loss of the respondents' privacy induced by these parameters. The following ratios λ_1 and λ_0 of conditional probabilities may be used for this purpose (*cf.* for example the similar "measures of jeopardy" in Leysieffer and Warner 1976, page 650):

$$\lambda_j = \frac{\max[P(y_i = j \mid i \in U_A),\ P(y_i = j \mid i \in U_A^c)]}{\min[P(y_i = j \mid i \in U_A),\ P(y_i = j \mid i \in U_A^c)]} \quad (10)$$

$(1 \le \lambda_j \le \infty;\ j = 1,\ 0)$.

For $j = 1$ (10) refers to the privacy protection with respect to a "yes"-, for $j = 0$ with respect to a "no"-answer. For the standardized questioning design these "λ-measures" of loss of privacy are given by

$$\lambda_1 = \frac{\max[a + b;\ b]}{\min[a + b;\ b]} \quad (11)$$

and

$$\lambda_0 = \frac{\max[1 - (a + b);\ 1 - b]}{\min[1 - (a + b);\ 1 - b]}. \quad (12)$$

$\lambda_1 = \lambda_0 = 1$ indicates a totally protected privacy. This means that the answer of the responding unit contains absolutely no information on the subject under study. This applies for $a = 0$. The more the λ-measures differ from unity, the more information about the characteristic under study is contained in the answer on the record. At the same time the efficiency of the estimation increases (see below), but the individual's protection against the data collector decreases. For the direct questioning design with $p_1 = 1$, where no masking of the variable is done at all, these measures are given by $\lambda_1 = \lambda_0 = \infty$.

Let the values $\lambda_{1,\,\mathrm{opt}}$ and $\lambda_{0,\,\mathrm{opt}}$ be the maximum λ-values of (11) and (12), that the agency considers to allow enough disclosure protection for the records. In the case of the strategy's usage as to avoid nonresponse and untruthful answering in surveys we may also model the respondents' willingness to cooperate as a function of perceived privacy protection. If the privacy of the respondents is sufficiently protected by the randomization device their full cooperation is assumed. Exceeding the limits $\lambda_{1,\,\mathrm{opt}}$ and/or $\lambda_{0,\,\mathrm{opt}}$ would then automatically introduce untruthful answering and nonresponse into the survey and therefore set us back to the starting point of the problem. Fidler and Kleinknecht (1977) showed in their study for design $ST11$ (Table 1) containing nine variables of very different levels of sensitivity, that their choice of the design parameters ($p_1 = 10/16$, $p_4 = p_5 = 3/16$) yielded nearly full and truthful response for each variable including sexual behaviour (ibd., page 1048). Inserting these values in (11) and (12) gives $\lambda_1 = \lambda_0 = 13/3$. This finding corresponds in the main with results that can be derived from the experiment by Soeken and Macready (1982) and with recommendations given by Greenberg *et al.* (1969). Therefore choosing $\lambda_{1,\,\mathrm{opt}}$ and/or $\lambda_{0,\,\mathrm{opt}}$ close to a value of 4 could be a good choice for most variables, when the standardized randomized response method is used to avoid refusals and untruthful answering of respondents in a survey.

Without loss of generality let us assume subsequently, that we will choose the two categories of the variable under study in such way, that the membership of U_A is at least as sensitive as the membership of U_A^c $(1 \le \lambda_{1,\,\mathrm{opt}} \le \lambda_{0,\,\mathrm{opt}} \le \infty)$. From (11) and (12) the terms a and b can be expressed by the λ-values λ_1 and λ_0. Their sum is given by:

$$a + b = \frac{1 - \dfrac{1}{\lambda_0}}{1 - \dfrac{1}{\lambda_1 \cdot \lambda_0}} \quad (13)$$

with

$$b = \frac{\dfrac{1}{\lambda_1} \cdot \left(1 - \dfrac{1}{\lambda_0}\right)}{1 - \dfrac{1}{\lambda_1 \cdot \lambda_0}} \quad (14)$$

and

$$a = \frac{\left(1 - \frac{1}{\lambda_1}\right) \cdot \left(1 - \frac{1}{\lambda_0}\right)}{1 - \frac{1}{\lambda_1 \cdot \lambda_0}}. \qquad (15)$$

We keep the double ratios on the right of (14) and (15) to find easily the limits for $\lambda_1 \to \infty$ and $\lambda_0 \to \infty$ respectively.

This means that for a given sampling design P the extent of the term $(b \cdot (1-b)/a^2) \cdot \sum_U (1/\pi_i) + (1 - 2 \cdot b - a/a) \cdot \sum_U (x_i/\pi_i)$ in the variance expression (6) does not depend on a single value of the design parameters, but on their aggregated effect on the loss of privacy measured by λ_1 and λ_0. Questioning designs with the same λ-values are equally efficient. Designs with larger λ_1 and/or λ_0 are less efficient than designs with lower λ's.

4. Optimum questioning designs

It does depend on the type of re-identification risk or sensitivity of the subject under study which of the special cases of the standardized randomized response strategy of Table 1 can be most efficient for given λ-measures. Strategies $ST5$ and $ST8$ can never perform best, because they do always protect a "no"-answer more than a "yes".

For a nonidentifying (or nonsensitive) variable (like for instance the season of birth), where $\lambda_{1,\,\mathrm{opt}} = \lambda_{0,\,\mathrm{opt}} = \infty$ applies, only the direct questioning design ($ST1$ of Table 1) can achieve the variance-optimum performance (see Table 2, which shows these values of the design parameters, which guarantee the best performance of the estimator $\hat{\pi}_A$; to be able to use Table 2 properly the categorial variable under study has to be classified according to the following categories: C_1: The variable is not sensitive at all ($\lambda_{1,\,\mathrm{opt}} = \lambda_{0,\,\mathrm{opt}} = \infty$); C_2: Only the membership of group U_A is sensitive, but not of U_A^c ($\lambda_{1,\,\mathrm{opt}} < \lambda_{0,\,\mathrm{opt}} = \infty$); C_3: The membership of both groups U_A and U_A^c is sensitive, but not equally ($\lambda_{1,\,\mathrm{opt}} < \lambda_{0,\,\mathrm{opt}} < \infty$); C_4: The membership of U_A and of U_A^c is equally sensitive ($\lambda_{1,\,\mathrm{opt}} = \lambda_{0,\,\mathrm{opt}} < \infty$), which shows these values of the design parameters, which guarantee the best performance of the estimator $\hat{\pi}_A$). Although the other designs can be used for such variables, they do unnecessarily protect the privacy of the respondents in some way. This has to be paid by a loss of accuracy of the estimation of π_A. But for $p_1 = 1$ ($a = 1$ and $b = 0$) the variance of $\hat{\pi}_A$ (5) turns to the common formula of the direct questioning with the assumption of full response: $V_P(\hat{\pi}_A) = 1/N^2 \cdot V_P(\sum_s x_i/\pi_i)$.

For a variable, of which only the membership of U_A, but not of U_A^c is sensitive (for instance: U_A = set of drug users within the last year; $U_A^c = U - U_A$) there is $\lambda_{1,\,\mathrm{opt}} < \lambda_{0,\,\mathrm{opt}} = \infty$. Calculating (14) and (15) for $1 < \lambda_1 < \infty$ and $\lambda_0 \to \infty$ gives $a = 1 - b$ and inserting this into (6) leads to the following expression for the variance of the estimator:

$$V_P(\hat{\pi}_A) = \frac{1}{N^2} \cdot \left[V_P\left(\sum_s \frac{x_i}{\pi_i} \right) + \frac{b}{1-b} \cdot \left(\sum_U \frac{1}{\pi_i} - \sum_U \frac{x_i}{\pi_i} \right) \right]. \qquad (16)$$

Looking for those values of the design parameters, for which the standardized randomized response strategy can achieve this variance and for which equations (14) to (15) hold, we do find that in this case there is only one solution! The only questioning design, that is able to perform optimally, is $ST4$. Its variance-optimum design parameters are given by $p_1 = (\lambda_1 - 1)/\lambda_1$ and $p_4 = 1 - p_1$ (see Table 2). This means, that with probability $p_1 = (\lambda_1 - 1)/\lambda_1$ a respondent is asked the question on membership of U_A and with the remaining probability he or she is instructed to say "yes". In this way the data collector is only able to conclude from a "no"-answer directly on the nonsensitive *non*-possession of A but not from a "yes"-answer on the possession of this sensitive or identifying attribute.

Questioning design $ST1$ is not applicable for such subjects, because it does not protect the respondent's privacy in case of a "yes"-answer at all. All the other procedures protect a "no"-answer more than necessary. Therefore they may be used, but they cannot achieve the efficiency of $ST4$.

If the membership of both U_A and U_A^c is sensitive, so that the variable is sensitive as a whole (for instance: U_A = set of married people, who had at least one sexual intercourse with their partners last week; $U_A^c = U - U_A$), $\lambda_{1,\,\mathrm{opt}} \le \lambda_{0,\,\mathrm{opt}} < \infty$ applies. In this case neither the direct questioning on the subject nor design $ST4$ can be used because they are not able to protect both possible answers.

The other designs are applicable for such topics, but Warner's design cannot achieve the efficiency of the others, if $\lambda_{1,\,\mathrm{opt}} < \lambda_{0,\,\mathrm{opt}}$. The reason is that this design always protects the respondent's privacy with respect to a "yes"-answer equally to a "no"-answer. But if $\lambda_{1,\,\mathrm{opt}} = \lambda_{0,\,\mathrm{opt}}$ despite to the claims of some publications in the past (see for instance: Greenberg *et al.* 1969, page 526f, Mangat and Singh 1990, page 440, Singh *et al.* 2003, page 518f) there is *not one* randomized response technique that can perform *better* than Warner's technique $ST2$ with the optimum design parameters p_1 and p_2 according to Table 2. For $ST7$ this is only valid for $\lambda_{1,\,\mathrm{opt}} < \lambda_{0,\,\mathrm{opt}}$. Therefore $ST7$ is the perfect supplement of $ST2$, for which the very opposite is true.

Table 2
Optimum design parameters for given λ_1 and λ_0 and different types of sensitivity of the variable under study

Questioning design (Subject category)	Variance-optimum design parameters
$ST1 \ (C_1)$	$p_1 = 1$
$ST2 \ (C_4)$	$p_1 = \frac{\lambda_1}{\lambda_1+1},\ p_2 = 1 - p_1$
$ST3 \ (C_3, C_4)$	$\pi_B = \frac{\lambda_0-1}{\lambda_1+\lambda_0-2},\ p_1 = \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},\ p_3 = 1 - p_1$
$ST4 \ (C_2)$	$p_1 = \frac{\lambda_1-1}{\lambda_1},\ p_4 = 1 - p_1$
$ST6 \ (C_4)$	$\pi_B = 0.5,\ p_1\text{: } \frac{\lambda_1-1}{\lambda_1+1} < p_1 < \frac{\lambda_1}{\lambda_1+1},\ p_2 = p_1 - \frac{\lambda_1-1}{\lambda_1+1},$ $p_3 = 1 - p_1 - p_2$
$ST6 \ (C_3)$	$\pi_B\text{: } \frac{\lambda_0-1}{\lambda_1+\lambda_0-2} < \pi_B < 1,$ $p_1 = \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1} + \frac{(\lambda_1-1)\cdot\pi_B-(\lambda_0-1)\cdot(1-\pi_B)}{(\lambda_1\cdot\lambda_0-1)\cdot(2\pi_B-1)},$ $p_2 = p_1 - \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},\ p_3 = 1 - p_1 - p_2$
$ST7 \ (C_3)$	$p_1 = \frac{\lambda_1\cdot\lambda_0-\lambda_0}{\lambda_1\cdot\lambda_0-1},\ p_2 = \frac{\lambda_1-1}{\lambda_1\cdot\lambda_0-1},\ p_4 = 1 - p_1 - p_2$
$ST9 \ (C_3, C_4)$	$\pi_B\text{: } 0 < \pi_B < \frac{\lambda_0-1}{\lambda_1+\lambda_0-2},\ p_1 = \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},$ $p_3 = \frac{\lambda_1-1}{(\lambda_1\cdot\lambda_0-1)\cdot(1-\pi_B)},\ p_4 = 1 - p_1 - p_3$
$ST10 \ (C_3, C_4)$	$\pi_B\text{: } \frac{\lambda_0-1}{\lambda_1+\lambda_0-2} < \pi_B < 1,\ p_1 = \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},$ $p_3 = \frac{\lambda_0-1}{(\lambda_1\cdot\lambda_0-1)\cdot\pi_B},\ p_5 = 1 - p_1 - p_3$
$ST11 \ (C_3, C_4)$	$p_1 = \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},\ p_4 = \frac{\lambda_0-1}{\lambda_1\cdot\lambda_0-1},\ p_5 = 1 - p_1 - p_4$
$ST12 \ (C_3, C_4)$	$p_1\text{: } \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1} < p_1 < \frac{\lambda_1\cdot\lambda_0-\lambda_0}{\lambda_1\cdot\lambda_0-1},\ p_2 = p_1 - \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},$ $\pi_B\text{: } 0 < \pi_B < \frac{\lambda_0-1-p_2\cdot(\lambda_1\cdot\lambda_0-1)}{\lambda_1+\lambda_0-2-2p_2\cdot(\lambda_1\cdot\lambda_0-1)},\ p_3 = \frac{\lambda_1-1-p_2\cdot(\lambda_1\cdot\lambda_0-1)}{(\lambda_1\cdot\lambda_0-1)\cdot(1-\pi_B)},$ $p_4 = 1 - \sum_{i=1}^{3} p_i$
$ST13 \ (C_3, C_4)$	$p_1\text{: } \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1} < p_1 < \frac{\lambda_1\cdot\lambda_0-\lambda_0}{\lambda_1\cdot\lambda_0-1},\ p_2 = p_1 - \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},$ $\pi_B\text{: } \frac{\lambda_0-1-p_2\cdot(\lambda_1\cdot\lambda_0-1)}{\lambda_1+\lambda_0-2-2p_2\cdot(\lambda_1\cdot\lambda_0-1)} < \pi_B < 1,\ p_3 = \frac{\lambda_0-1-p_2\cdot(\lambda_1\cdot\lambda_0-1)}{(\lambda_1\cdot\lambda_0-1)\cdot\pi_B},$ $p_5 = 1 - \sum_{i=1}^{3} p_i$
$ST14 \ (C_3, C_4)$	$p_1\text{: } \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1} < p_1 < \frac{\lambda_1\cdot\lambda_0-\lambda_0}{\lambda_1\cdot\lambda_0-1},\ p_2 = p_1 - \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},$ $p_4 = \frac{\lambda_0-1}{\lambda_1\cdot\lambda_0-1} - p_2,\ p_5 = 1 - p_1 - p_2 - p_4$
$ST15 \ (C_3, C_4)$	$\pi_B\text{: } 0 < \pi_B < 1,\ p_1 = \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},$ $p_3\text{: } 0 < p_3 < \frac{\lambda_1-1}{(\lambda_1\cdot\lambda_0-1)\cdot(1-\pi_B)},\ p_4 = \frac{\lambda_0-1}{\lambda_1\cdot\lambda_0-1} - p_3\cdot\pi_B,$ $p_5 = 1 - p_1 - p_3 - p_4$
$ST16 \ (C_3, C_4)$	$\pi_B\text{: } 0 < \pi_B < 1,\ p_1\text{: } \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1} < p_1 < \frac{\lambda_1\cdot\lambda_0-\lambda_0}{\lambda_1\cdot\lambda_0-1},$ $p_2 = p_1 - \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},\ p_3\text{: } 0 < p_3 < \frac{\lambda_1\cdot\lambda_0-\lambda_0}{\lambda_1\cdot\lambda_0-1} - p_1,$ $p_4 = \frac{\lambda_0-1}{\lambda_1\cdot\lambda_0-1} - p_2 - p_3\cdot\pi_B,\ p_5 = 1 - \sum_{i=1}^{4} p_i$

All others of the designs of Table 1 like $ST11$ or $ST14$ can perform equally efficient for $\lambda_{1,\,\mathrm{opt}} \le \lambda_{0,\,\mathrm{opt}} < \infty$, if the design parameters are chosen according to the restrictions (14) to (15). Among them Greenberg *et al.*'s strategy with known π_B $(ST3)$ has on the one hand the advantage over Warner's design to be able to perform optimally also if $\lambda_{1,\,\mathrm{opt}} < \lambda_{0,\,\mathrm{opt}}$. On the other hand, however, it has the disadvantage (like $ST6$), that the size π_B of subpopulation U_B is completely predetermined (or at least bounded by an interval), if we want to achieve the optimum efficiency. This means in practice, that we have to find a subpopulation not related to the possession and nonpossession of attribute A *and* of appropriate relative size to be able to achieve the estimator's optimum accuracy. In principle this also applies to $ST9$, $ST10$, $ST12$ and $ST13$, but looking at the presettings of design parameter π_B, it turns out that $ST9$ and $ST10$ as well as $ST12$ and $ST13$ perfectly complement each other so that in fact any subset U_B of the population can be used. Finally the most complex special cases, $ST15$ and $ST16$, of our standardized randomized response strategy can both be used with any subpopulation $U_B \subset U$ to achieve the best performance.

5. A real-data example

An empirical study was carried out to demonstrate the applicability of the strategy as a questioning design. For this purpose the population of 80 students, who attended the author's course on "Statistics II" at the Johannes Kepler University in Linz (Austria) during the spring term of 2009, volunteered for the survey. The subject under study was academic cheating behaviour. To this end cheating was defined as any behaviour, that was not allowed in the written exams (including just looking at the test scripts of other students or the use of forbidden documents). It is beyond doubt that this subject is sensitive for such a population. Moreover during the survey all of the students were sitting in one lecture room. The parameter of interest was the proportion of the population of students, that fudged on at least one of the exams of the previous semester (including the exam of the author's course on "Statistics I"). Therefore it is beyond reasonable doubt to assume, that direct questioning on the subject would have resulted into a substantial underestimation of this proportion. An empirical study of Scheers and Dayton (1987) for instance showed very small proportions for almost all different cheating behaviours asked, when the subject in question was asked directly. The use of Greenberg's randomized response strategy $ST3$ lead to a significant increase of these proportions (ibd., page 68).

Apparently, for the variable of interest the membership of group U_A, formed by the "cheaters", is sensitive, but not the membership of the complementary set U_A^c. Therefore in accordance with the recommendations of Section 4 we decided to use questioning design $ST4$ for our survey and to compare it with Warner's strategy $ST2$. The λ-values of loss of privacy were fixed at $\lambda_1 = 4$ and $\lambda_0 = \infty$. From Table 2 we calculated $p_1 = 0.75$ and $p_4 = 0.25$ as the variance-optimum design parameters of $ST4$. To achieve these probabilities the students were asked to throw two dice without showing the result to somebody else and answer in a questionnaire the question "Did you cheat at the exams at least one time?" only if the sum of the numbers on the dice was 5 to 10. Otherwise they should just respond "yes".

Previous to the survey some effort was made to explain the consequences of this randomization strategy on the privacy protection. After giving the answer on the first sheet of the questionnaire, only these sheets were collected. 63 out of the 80 persons answered "yes". 20 of 80 students were expected to do so, because they received the "say yes-instruction". Therefore expected 43 of 60 other students should have answered "yes" on the sensitive question. The estimator for π_A is given by

$$\hat{\pi}_A^{ST4} = \frac{\hat{\pi}_y^{ST4} - p_4}{p_1} = \frac{0.7875 - 0.25}{0.75} = 0.71\dot{6}.$$

For this population survey the estimated variance of $\hat{\pi}_A$ is then

$$\hat{V}(\hat{\pi}_A^{ST4}) = \frac{1 - p_1}{n \cdot p_1} \cdot (1 - \hat{\pi}_A^{ST4}) = 1.181 \cdot 10^{-3}.$$

After this questioning design was completed, the students were asked directly on the second sheet of the questionnaire, whether they had truthfully answered the first question or not. Only four students said that this was not the case. This means, that – if that's true – it is likely that 4 more students did actually cheat. The next question to answer was, if they would still cooperate, if p_1 (of $ST4$) would be higher than 0.75. 32 of 80 students agreed to do so, but the others did not. Obviously (at least) four of them did not cooperate when p_1 was 0.75.

Finally, Warner's technique was applied with the same sensitive question as $ST4$ before. To come close to a λ_1-level of 4 – indicating the same loss of privacy as to a "yes"-answer for both questioning designs –, the sum of the numbers of two dice had to be 3 to 9 to apply a design parameter $p_1 = 0.80\dot{5}$. The λ-measures of loss of privacy for this choice are given by $\lambda_1 = \lambda_0 = 4.143$, indicating a slightly higher loss of privacy compared to $ST4$. With a probability of $0.80\dot{5}$ the students had to answer "Are you a member of U_A? " and with the remaining probability the alternative "Are you a member of U_A^c? ".

Now only 38 of 80 persons gave a "yes"-answer. This results in an estimated proportion of "cheaters" of

$$\hat{\pi}_A^{ST2} = \frac{\hat{\pi}_y^{ST2} - p_2}{p_1 - p_2} = \frac{0.475 - 0.19\dot{4}}{0.6\dot{1}} = 0.45\dot{9}0.$$

Additionally to the slight increase of the objective loss of privacy there is another reasonable explanation for this significantly lower result. Although λ_1 did not change that much, some test persons must have been irritated by the raise of p_1 up to $0.80\dot{5}$ after being asked for $ST4$, if they would still cooperate, if p_1 would be higher than 0.75. Not being able to distinguish between the loss of privacy caused by different design parameters in different questioning designs, some of the "cheaters" did not want to answer truthfully again. Just to demonstrate the effect of the different questioning designs on the efficiency of the estimation process we calculate the estimator of the variance of $\hat{\pi}_A^{ST2}$:

$$\hat{V}(\hat{\pi}_A^{ST2}) = \frac{p_1 \cdot (1 - p_1)}{n \cdot (2p_1 - 1)^2} = 5.243 \cdot 10^{-3}.$$

The reason for this considerable increase of the estimated variance is, that Warner's strategy does protect a "no"-answer always in the same way as a "yes". Since in our case a "no"-answer does not have to be protected at all, this unnecessary protection has to be paid in terms of accuracy.

6. Summary

Randomized response strategies have originally been developed to reduce the nonreponse as well as the untruthful answering rate for sensitive subjects in sample surveys, but they can be applied as masking techniques for public use microdata files as well. The standardization of these techniques for the estimation of proportions developed in this paper provides an opportunity to derive a general formula for the variance of the estimator under probability sampling. Different questioning designs, partly published, partly – to the best of our knowledge – unpublished up to now, can be regarded as special cases of the standardized strategy (see Table 1). For the purpose of a comparison of the accuracy of these designs it is essential to include the levels of privacy protection offered by them in our considerations. Doing this by means of the "λ-measures" of loss of privacy explicated in Section 3 a completely new picture has to be painted in comparison to almost all publications in the past as far as the author knows them. It turns out that the identifying or sensitive subjects have to be classified into different categories in order to find the variance-minimum questioning designs for a given privacy protection (see Table 2). The first category consists of subjects, which are not sensitive at all. The second comprises topics, where only the possession but not the nonpossession of a certain attribute is embarrassing to the respondents. The last category is formed by subjects, which are sensitive as a whole.

For subjects out of the first category it is clear enough that no strategy can be more efficient than the direct questioning on the subject ($ST1$ of Table 1).

Concerning topics of the second category there is just one design available, that can achieve the minimum variance of the estimator. This is the questioning design in which each respondent either with probability p_1 has to answer the question on membership of the sensitive group or with probability $1 - p_1$ is instructed to answer "yes" ($ST4$). All the other special cases of the standardized strategy protect the interviewee's privacy not only in case of a "yes"-answer like $ST4$ does, but also in case of a "no"-answer. Therefore their performances cannot reach the minimum achievable level.

For subjects out of the third category it is shown, that contrary to the claim of other publications, there is not one single strategy available that can perform better than Warner's of 1965 as long as the membership of the subgroup under investigation is equally sensitive to the membership of its complement. A lot of other designs are *equally* efficient as Warner's but not a single one is *more* efficient.

For the variables of this category, where the membership of one group is sensitive, but not equally sensitive as the membership of the complementary one, the situation changes dramatically: Compared under the same levels of privacy protection Warner's technique is not able to achieve the best achievable performance of the standardized randomized design anymore, whereas many other strategies can. For some of the designs including the question on membership of a nonsensitive subpopulation not related to the attribute under study, it is required to find an adequate subpopulation of predetermined relative size. Other designs can be used with subpopulations of any size and are therefore more practicable. Therefore a data collector or publisher could select that one of the equally efficient designs, that seems to be more easily applicable than the others.

Acknowledgements

The author is very grateful to the Associate Editor and two referees for their valuable comments and suggestions.

Appendix

Proofs of theorems 1 and 2

Proof of Theorem 1:

$$E(\hat{\pi}_A) = \frac{1}{N} \cdot E_P\left(E_R\left(\sum_s \frac{\hat{x}_i}{\pi_i} \mid s\right)\right)$$
$$= \frac{1}{N} \cdot E_P\left(\sum_s \frac{x_i}{\pi_i}\right) = \frac{1}{N} \cdot \sum_U x_i = \pi_A.$$

The variance of estimator (5) is given by

$$V(\hat{\pi}_A) = V_P(E_R(\hat{\pi}_A \mid s)) + E_P(V_R(\hat{\pi}_A \mid s)).$$

Then

$$V_P(E_R(\hat{\pi}_A \mid s)) = \frac{1}{N^2} \cdot V_P\left(\sum_s \frac{x_i}{\pi_i}\right).$$

Let the sample inclusion indicator

$$I_i = \begin{cases} 1 & \text{if unit } i \in s, \\ 0 & \text{otherwise.} \end{cases}$$

Because the covariance $C_R(\hat{x}_i, \hat{x}_j \mid s) = 0 \ \forall \ i \neq j$, for the second summand of $V(\hat{\pi}_A)$ applies

$$E_P(V_R(\hat{\pi}_A \mid s)) = E_P\left(\frac{1}{N^2} \cdot V_R\left(\sum_U I_i \cdot \frac{\hat{x}_i}{\pi_i} \mid s\right)\right)$$
$$= E_P\left(\frac{1}{N^2} \cdot \sum_U \frac{I_i^2}{\pi_i^2} \cdot V_R(\hat{x}_i)\right)$$
$$= \frac{1}{N^2} \cdot \sum_U \frac{V_R(\hat{x}_i)}{\pi_i}.$$

For $V_R(\hat{x}_i)$ we have

$$V_R(\hat{x}_i) = \frac{1}{a^2} \cdot V_R(y_i)$$

and

$$V_R(y_i) = b + a \cdot x_i - (b + a \cdot x_i)^2$$
$$= (b + a \cdot x_i) \cdot (1 - b - a \cdot x_i)$$
$$= b \cdot (1 - b) + a \cdot (1 - 2 \cdot b - a) \cdot x_i.$$

Then

$$E_P(V_R(\hat{\pi}_A \mid s)) =$$
$$\frac{1}{N^2} \cdot \left(\frac{b \cdot (1 - b)}{a^2} \cdot \sum_U \frac{1}{\pi_i} + \frac{1 - 2 \cdot b - a}{a} \cdot \sum_U \frac{x_i}{\pi_i}\right).$$

This completes the proof of Theorem 2.

References

Chaudhuri, A., and Mukerjee, R. (1987). *Randomized Response*. New York: Marcel Dekker.

Dalenius, T. (1977). Privacy transformations for statistical information systems. *Journal of Statistical Planning and Inference*, 1, 73-86.

Dalenius, T., and Reiss, S.P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6, 73-85.

Defays, D., and Anwar, M.N. (1998). Masking microdata using micro-aggregation. *Journal of Official Statistics*, 14 (4), 449-461.

Domingo-Ferrer, J., and Mateo-Sanz, J.M. (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14 (1), 189-201.

Fidler, D.S., and Kleinknecht, R.E. (1977). Randomized response versus direct questioning: Two data collection methods for sensitive information. *Psychological Bulletin*, 84 (5), 1045-1049.

Fuller, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9 (2), 383-406.

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J. and de Wolf, P.-P. (1998). Post randomisation for statistical disclosure control: Theory and implementation. *Journal of Official Statistics*, 14 (4), 463-478.

Greenberg, B.G., Abul-Ela, A.-L.A., Simmons, W.R. and Horvitz, D.G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, 64, 520-539.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R. (2004). *Survey Methodology*. Hoboken: John Wiley & Sons, Inc.

Horvitz, D.G., Shah, B.V. and Simmons, W.R. (1967). The unrelated question randomized response model. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 65-72.

Hua, M., and Pei, J. (2008). A survey of utility-based privacy-preserving data transformation methods. In: *Privacy-preserving Data Mining*: *Models and Algorithms*, (Eds., C.C. Aggarwal and P.S. Yu), New York: Springer, 207-238.

Kim, J. (1987). A further development of the randomized response technique for masking dichotomous variables. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 239-244.

Kim, J.M., and Warde, W.D. (2005). A mixed randomized response model. *Journal of Statistical Planning and Inference*, 133, 211-221.

Kuk, A.Y.C. (1990). Asking sensitive questions indirectly. *Biometrika*, 77 (2), 436-438.

Landsheer, J.A., van der Heijden, P. and van Gils, G. (1999). Trust and understanding, two psychological aspects of randomized response. *Quality and Quantity*, 33, 1-12.

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

Mangat, N.S. (1992). Two stage randomized response sampling procedure using unrelated question. *Journal of the Indian Society of Agricultural Statistics*, 44, 82-87.

Mangat, N.S. (1994). An improved randomized response strategy. *Journal of the Royal Statistical Society*, Series B, 56, 93-95.

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

Mangat, N.S., Singh, S. and Singh, R. (1993). On the use of a modified randomization device in randomized response inquiries. *Metron*, 51, 211-216.

Nathan, G. (1988). A bibliography of randomized response: 1965-1987. *Survey Methodology*, 14, 331-346.

Quatember, A. (2007). Comparing the efficiency of randomized response techniques under uniform conditions. *IFAS Research Paper Series*, 23, www.ifas.jku.at/e2550/e2756/index_ger.html.

Rosenberg, M.J. (1980). Categorical data analysis by a randomized response technique for statistical disclosure control. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 311-316.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.

Scheers, N.J., and Dayton, C.M. (1987). Improved estimation of academic cheating behaviour using the randomized response technique. *Research in Higher Education*, 26 (1), 61-69.

Singer, E., Mathiowetz, N.A. and Couper, M.P. (1993). The impact of privacy and confidentiality concerns on survey participation: The case of the 1990 U.S. Census. *The Public Opinion Quarterly*, 57 (4), 465-482.

Singer, E., van Hoewyk, J. and Neugebauer, R.J. (2003). Attitudes and behavior: The impact of privacy and confidentiality concerns on participation in the 2000 Census. *The Public Opinion Quarterly*, 67 (3), 368-384.

Singh, R., Singh, S., Mangat, N.S. and Tracy, D.S. (1995). An improved two stage randomized response strategy. *Statistical Papers*, 36, 265-271.

Singh, S., Horn, S., Singh, R. and Mangat, N.S. (2003). On the use of modified randomization device for estimating the prevalence of a sensitive attribute. *Statistics in Transition*, 6 (4), 515-522.

Singh, S., Singh, R., Mangat, N.S. and Tracy, D.S. (1994). An alternative device for randomized responses. *Statistica*, 54, 233-243.

Skinner, C., Marsh, C., Openshaw, S. and Wymer, C. (1994). Disclosure control for census microdata. *Journal of Official Statistics*, 10 (1), 31-51.

Soeken, K.L., and Macready, G.B. (1982). Respondents' perceived protection when using randomized response. *Psychological Bulletin*, 92 (2), 487-489.

Tracy, D.S., and Mangat, N.S. (1996). Some developments in randomized response sampling during the last decade – A follow up of review by Chaudhuri and Mukerjee. *Journal of Applied Statistical Science*, 4 (2/3), 147-158.

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 (2), 269-288.

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

Warner, S.L. (1971). The linear randomized response model. *Journal of the American Statistical Association*, 66, 884-888.

Willenborg, L., and de Waal, T. (1996). *Statistical Disclosure Control in Practice*. New York: Springer.

Winkler, W.E. (2004). Masking and re-identification methods for public-use microdata: Overview and research problems. *Research Report Series of the Statistical Research Division of the U.S. Bureau of the Census*, #2004-06.