

Selected Scrambled Response Models

Giancarlo Diana and Pier Francesco Perri

Abstract In this paper we discuss the problem of obtaining reliable data on a sensitive quantitative variable without jeopardizing respondents privacy. The collected information is obtained by asking the respondents to perturb the response through a scrambling mechanism. A general device allowing for the use of an auxiliary variable is illustrated as well as a class of estimators for the unknown mean of a sensitive variable. A number of scrambled response models are shown and others discussed in terms of the efficiency of the estimates and the privacy protection guaranteed to respondents.

Key words: Sample Design and Estimation, Privacy Protection, Sensitive Variable, Class of Estimators.

1 Introduction

Estimates derived from sample surveys are subject to two main sources of error: *sampling* and *nonsampling error*. The former is due to the variation that is present because a sample is studied instead of the whole population. Sampling errors can be remediated by employing different strategies which increase the representativeness of the sample and improve the efficiency of the estimates. The latter is much more problematic to treat because it can stem from many sources such as nonresponse,

Giancarlo Diana

Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121 Padova, Italy. e-mail: giancarlo.diana@unipd.it

Pier Francesco Perri

Department of Economics and Statistics, University of Calabria, Cubo OC, 87036 Arcavacata di Rende (CS), Italy. e-mail: pierfrancesco.perri@unical.it

incorrect responses, falsification of information, coding and computer processing errors, inaccuracies in the sampling frame, reporting errors and others.

Nonsampling errors become particularly relevant in surveys where people are asked about private and confidential topics such as gambling, alcoholism, sexual behavior, abortion, drug taking, tax evasion, illegal income, and so on. In fact, interviewees often refuse to answer or give untruthful or misleading responses to questions they may find embarrassing or harmful in some way. Even when the interviewers do their best to guarantee confidentiality, subjects can be skeptical and may be reluctant to supply truthful replies. Obtaining valid and reliable information on sensitive issues depends on the cooperation of the respondents, and this depends on confidentiality of the responses.

To reduce the rate of non-response and minimize the underreporting of embarrassing, threatening, or even incriminating behaviors, survey statisticians have developed a number of procedures which ensure interviewee anonymity or, at least, a high degree of confidentiality. When sensitive topics are surveyed, researchers often use self-administered questionnaires with paper and pencil, computer assisted self-interviews or Internet-delivered interviews. An entirely different approach for eliciting information on sensitive data while ensuring respondents' privacy protection is the ingenious interviewing procedure pioneered by Warner (1965) and known as the randomized response (RR) technique. The rationale of the technique is simple: respondents are instructed to give partially misclassified (or perturbed) responses by using a randomization device, such as dice or a deck of cards. This can be done, for example, by asking the respondents to reply honestly to the sensitive question or according to the outcome of the randomizing device. In this way, the interviewer does not know whether the answer refers to the sensitive question or is forced by the outcome of the device. Because the interviewer is unaware of the result of the device, the use of the RR procedure ensures that respondents cannot be identified on the basis of their answers.

Different studies have assessed the validity of RR methods and shown that these can significantly outperform conventional data collection methods and procure more honest and reliable answers (see, e.g., van der Heijden et al. (2000), Lara et al. (2004), Lensvelt-Mulders et al. (2005)). In the Netherlands, there is a well-established tradition towards the use of RR data collection in official surveys mainly concerned with measuring the prevalence of fraud, especially in the area of disability benefits (see Lensvelt-Mulders et al. (2006) for a brief excursus on the research). However, many authors use RR techniques in a wide range of empirical studies. Among others, Lara et al. (2004, 2006) have recently estimated the prevalence of abortion in Mexico; van der Heijden and Böckenholt (2008) discussed applications of the methodology in e-commerce; Ostapczuk et al. (2009) and Krumpal (2010) treated the issue of xenophobia and anti-semitism in Germany, while Arnab and Singh (2010) measured the impact of HIV/AIDS infection in Botswana.

Standard RR methods are used primarily in surveys which require a "yes" or "no" response to a sensitive question such as in the above-mentioned surveys, or a choice of responses from a set of nominal categories. The purpose is to estimate the proportion of people bearing a sensitive characteristic (see, e.g., Diana and Perri

(2009)). Nevertheless, there is a growing amount of theoretical literature dealing with situations where the response to a sensitive question results in a quantitative variable and the interest is the estimation of the mean or total of the sensitive variable. Consider, for instance, the problem of estimating the average personal income or the number of frauds and see Gjestvang and Singh (2009) for a recent application to the estimation of average GPA of students at St Cloud State University. In these cases, people are asked to algebraically scramble the true response by means of a coding mechanism.

Data perturbation is also of practical concern in the context of the *Statistical Disclosure Control* where information is masked - after data have been collected - to be released to third parties by data producers such as national statistical institutes (see, e.g. Willenborg and Waal (1996)).

In the remainder of this paper, we focus on some selected scrambled response models for quantitative data. We introduce a general approach which takes into account the possibility of using an auxiliary variable and we discuss a class of estimators for the mean of a sensitive variable of interest. The problem of evaluating the degree of privacy protection is also considered. In particular, in Section 2 we introduce a general mechanism to scramble the responses, and a class of estimators for the mean of the sensitive variable is given when the sample of respondents is selected according to simple random sampling with replacement (SRSWR). Finally, certain selected models to code the response are shown. In Section 3, two measures for assessing the degree of respondents' privacy protection are discussed and illustrated with reference to different simple scrambled models. Section 4 is devoted to comparing some models from the viewpoint of efficiency and privacy protection. Section 5 concludes the paper with some remarks.

2 A Class of Estimators for the Sensitive Mean

Let $\mathcal{U} = \{1, 2, \dots, N\}$ be a finite population of identifiable individuals and $Y \geq 0$ be a quantitative sensitive variable with mean and variance, $\mathbb{E}(Y) = \mu_y$ and $\mathbb{V}(Y) = \sigma_y^2$, assumed to be unknown.

In the framework of the RR methodology, many devices have been implemented to estimate μ_y . Most of these use a coding mechanism of the response on Y , i.e., respondents are asked to perturb the true value of Y algebraically through one or more random numbers generated from known scrambling distributions. For instance, Pollock and Bek (1976) considered the additive/multiplicative models which involve the respondents adding/multiplying the answer to the sensitive question by a random number from a known distribution. The multiplicative model was later taken up and investigated in depth by Eichhorn and Hayre (1983). Gupta et al. (2002) introduced an optional randomized response technique which is more efficient than Eichhorn and Hayre's scrambled device. Bar-Lev et al. (2004) generalized Eichhorn and Hayre's device by introducing a design parameter which is used for random-

izing the responses. Hereafter, we refer to coding procedures for quantitative as scrambled response (SR) models.

Most of the SR models in the literature do not consider the possibility of getting more accurate estimates of μ_y by using auxiliary variables directly at the estimation stage. Here we discuss this possibility when a non-sensitive auxiliary variable, say X , correlated to Y , is available. Suppose that μ_x and σ_x^2 are both known. Suppose that the values of Y [X] can be perturbed algebraically by means of positive scrambling variables W and U [T and H] whose distributions are known. The scrambling variables are assumed to be independent of each other and also independent of Y [X]. Finally, let $S = \varphi(W, U, Y)$ and $R = \psi(T, H, X)$ denote the coded responses on Y and X , respectively, according to the *scrambling functions* φ and ψ induced by the coding device.

In our approach, to estimate μ_y , a sample of n individuals is selected from the population and each respondent is asked to perform a Bernoulli trial with probability of success p . If this is successful, the respondent then gives the true values of both Y and X . In case of failure the values S and R are given. Obviously, the interviewer does not know the outcome of the Bernoulli experiment. Then, the distribution of the responses is

$$(Z, V) = \begin{cases} (Y, X), & \text{with probability } p \\ (S, R), & \text{with probability } 1 - p. \end{cases} \quad (1)$$

Under this generic scheme, we consider the following class of estimators of μ_y based on a SRSWR sample $\{(z_1, v_1), (z_2, v_2), \dots, (z_n, v_n)\}$ of n responses

$$\hat{\mu}_g = \frac{\bar{z}_d - c}{h}, \quad (h \neq 0) \quad (2)$$

where $\bar{z}_d = \bar{z} + b(\mu_v - \bar{v})$ is the *difference estimator* of μ_z with $\bar{z} = \sum_{i=1}^n z_i/n$, $\bar{v} = \sum_{i=1}^n v_i/n$, $\mathbb{E}(\bar{z}) = \mu_z$, $\mathbb{E}(\bar{v}) = \mu_v$; b , c and h are suitably chosen real constants such that $c + h\mu_y = \mathbb{E}(\bar{z})$. In particular, the values of c and h depend exclusively on the scrambling device adopted, whereas the choice of b is mainly linked to the efficient use of the auxiliary variable. The choice $b = 0$ excludes the possibility of using auxiliary information at the estimation stage and, thus, for fixed c and h , the class reduces to the specific estimator (with $\bar{z}_d = \bar{z}$) generated by the adopted SR method. Therefore, formulation (2) has the merit of including, under different coding procedures, both classical estimators without supplementary information and possible estimators based on an auxiliary variable.

It is straightforward to verify that $\hat{\mu}_g$ is unbiased for μ_y and has variance

$$\mathbb{V}(\hat{\mu}_g) = \frac{\mathbb{V}(\bar{z}_d)}{h^2} = \frac{1}{nh^2} (\sigma_z^2 - 2b\sigma_{vz} + b^2\sigma_v^2)$$

which attains its minimum bound

$$\mathbb{V}(\hat{\mu}_g)_{min} = \frac{\sigma_z^2}{nh^2} (1 - \rho_{vz}^2) \quad (3)$$

for $b_o = \sigma_{vz}/\sigma_v^2$, $\sigma_{..}$ and $\rho_{..}$ being the covariance and the correlation coefficient between the variables indicated in the subscript, respectively. This expression well emphasizes the role of the auxiliary variable in improving the accuracy of the estimates. In this regard, σ_z^2/nh^2 is the variance of the estimator which does not employ the auxiliary variable X and $(1 - \rho_{vz}^2)$ denotes the reduction in the variance due to the use of X through the traditional *regression estimator* (Cochran (1977)).

The optimum coefficient b_o is generally unknown and has to be estimated. It's consistent estimate is likely to be the least squares estimate

$$\hat{b}_o = \frac{s_{vz}}{s_v^2} = \frac{\sum_{i=1}^n (v_i - \bar{v})(z_i - \bar{z})}{\sum_{i=1}^n (v_i - \bar{v})^2}$$

which can be always calculated since v_i and z_i denote the scrambled responses provided by the i th sampled respondent. Note that this formulation avoids resorting to the sensitive variable whose values cannot be directly collected.

Remark. The extension of the results to the case where k auxiliary variables $X_1, \dots, X_k$ are used is straightforward. From Diana and Perri (2007), in (2) we can use $\bar{z}_d = \bar{z} + \sum_{j=1}^k \beta_{zv_j}(\mu_{v_j} - \bar{v}_j)$ where V_j is defined as in (1) according to the response on X_j and β_{zv_j} is the *partial regression coefficient* of Z on V_j . Therefore, the variance of the best estimator of Y becomes $\mathbb{V}(\hat{\mu}_g)_{min} = \frac{\sigma_z^2}{nh^2}(1 - \rho_{z.v_1 \dots v_k}^2)$, with $\rho_{z.v_1 \dots v_k}$ the *multiple correlation coefficient* of Z on $V_1, \dots, V_k$. Since $\rho_{z.v_1 \dots v_k}^2 \geq \rho_{vz}^2$, it turns out that the use of more than one auxiliary variable is certainly profitable.

2.1 Selected Scrambled Response Models

The class we have introduced in a general framework is extremely flexible since it can be easily adapted to the perturbing response mechanism that the researcher decides to use for the sensitive problem at hand. Once the choice of the model has been made, the coding functions φ and ψ assume an explicit form and the constant c and h are defined accordingly. Table 1 summarizes a number of examples concerning different specifications of the general SR model and gives the expressions for c and h to estimate μ_y from (2). Note that the choice $p = 0$ excludes any possibility of giving a direct response on Y , while for $p \in (0, 1)$ responses on Y can be partially perturbed.

To evaluate the performance of a particular scrambling scheme in terms of efficiency of the estimates we need to rewrite the following expressions

$$\mu_z = p\mu_y + (1-p)\mu_s, \quad \mu_v = p\mu_x + (1-p)\mu_r \quad (4)$$

$$\sigma_z^2 = p\mu_{2,y} + (1-p)\mu_{2,s} - \mu_z^2, \quad \sigma_v^2 = p\mu_{2,x} + (1-p)\mu_{2,r} - \mu_v^2 \quad (5)$$

$$\sigma_{vz} = p\sigma_{xy} + (1-p)\sigma_{rs} + p(1-p)(\mu_y - \mu_s)(\mu_x - \mu_r). \quad (6)$$

Table 1 Selected models from the generic scheme

ψ	ϕ	p	c	h
X	$Y+U$	$[0, 1]$	$(1-p)\mu_u$	1
X	YW	$[0, 1]$	0	$(1-p)\mu_w + p$
$X+H$	$Y+U$	$[0, 1]$	$(1-p)\mu_u$	1
XT	YW	$[0, 1]$	0	$(1-p)\mu_w + p$
X	$W(Y+U)$	$[0, 1]$	$(1-p)\mu_w\mu_u$	$p + (1-p)\mu_w$
X	$W[\alpha U + (1-\alpha)Y]$	$[0, 1]$	$(1-p)\alpha\mu_w\mu_u$	$(1-p)(1-\alpha)\mu_w + p$
X	$\xi(Y+U) + (1-\xi)WY$	$[0, 1]$	$(1-p)\xi\mu_u$	$(1-p)[\xi + (1-\xi)\mu_w] + p$
$T(X+H)$	$W(Y+U)$	$[0, 1]$	$(1-p)\mu_w\mu_u$	$p + (1-p)\mu_w$
$T[\alpha H + (1-\alpha)X]$	$W[\alpha U + (1-\alpha)Y]$	$[0, 1]$	$(1-p)\alpha\mu_w\mu_u$	$(1-p)(1-\alpha)\mu_w + p$
$\xi(X+H) + (1-\xi)TX$	$\xi(Y+U) + (1-\xi)WY$	$[0, 1]$	$(1-p)\xi\mu_u$	$(1-p)[\xi + (1-\xi)\mu_w] + p$

$\alpha, \xi \in [0, 1]$ are design parameters controlled by the experimenter

where $\mu_{2,\cdot}$ denotes the second moment of the variable indicated in the subscript. The details have been omitted since most of these expressions, as well as the guidelines for their derivation, are given in Diana and Perri (a,b).

3 Respondents' Privacy Protection

One relevant aspect, often underrated in RR theory, is the measure of respondents' privacy protection. An index used to measure privacy should indicate how closely the original values of the perturbed sensitive variable Y can be estimated. The closer these values, the higher the privacy disclosure. SR devices which ensure high performance in terms of efficiency are usually less protective. Therefore, a fair comparison of alternative scrambling strategies should necessarily take into account both efficiency and privacy protection considerations. Different ways of measuring confidentiality are discussed in the literature and a brief review is reported in Diana and Perri (a).

Starting from a recent proposal by Zaizai et al. (2009), we first consider the measure $\Delta = \mathbb{E}(Q - Y)^2 = \mathbb{E}_y \mathbb{E}_{sr}(Q - Y)^2$ where $\mathbb{E}_y$ and $\mathbb{E}_{sr}$ denote the expectations with respect to the study variable on the population and the scrambled responses, Q being an unbiased estimator, with respect the scrambling device, of the true value of Y . The magnitude of deviation of the unbiased coded response Q_i from the true undiscovered value Y_i determines the *degree of protection* to i th respondent induced by the adopted SR device. If $\Delta = 0$ then privacy is completely jeopardized, in other words the respondent's privacy has not been protected. Focusing, for simplicity, on scrambling methods allowing us to obtain only coded sensitive response (i.e., $p = 0$ and $Z = S$ in (1)), it follows that $Q = (S - c_0)/h_0$ where c_0 and h_0 denote the expressions for c and h when $p = 0$. Therefore, Δ measure can be expressed as

$$\Delta = \mathbb{E} \left(\frac{S - c_0}{h_0} - Y \right)^2 = \mathbb{E}_y \mathbb{E}_{sr} \left(\frac{S - c_0}{h_0} - Y \right)^2. \quad (7)$$

After some algebra, we get the form

$$\Delta = \frac{h_0^2 \mu_{2,y} - 2h_0 \mu_{ys} + 2c_0 h_0 \mu_y + \mu_{2,s} - 2c_0 \mu_s + c_0^2}{h_0^2} \quad (8)$$

which can be adapted to a given coding method by means of an appropriate specification of the constants c_0 and h_0 , as well as the moments of second order.

4 Comparisons

For illustrative purposes, we now investigate the following simple SR models which are taken as a starting point for the analysis of more complex devices

Author	Model	Coded response Z	Estimator
Pollock and Bek (1976)	additive	$Y + U$	$\hat{\mu}_{PB} = \bar{z} - \mu_u$
Eichhorn and Hayre(1983)	multiplicative	YW	$\hat{\mu}_{EH} = \frac{\bar{z}}{\mu_w}$
Saha (2007)	mixed1	$W(Y + U)$	$\hat{\mu}_1 = \frac{\bar{z}}{\mu_w} - \mu_u$
Diana and Perri (b)	mixed2	$W[\alpha U + (1 - \alpha)Y]$	$\hat{\mu}_2 = \frac{\bar{z} - \alpha \mu_w \mu_u}{(1 - \alpha) \mu_w}$
Diana and Perri (b)	mixed3	$\xi(Y + U) + (1 - \xi)WY$	$\hat{\mu}_3 = \frac{\bar{z} - \xi \mu_u}{\xi + (1 - \xi) \mu_w}$

We observe that these models do not involve the auxiliary variable X . Considerations on the performance of the models are detailed in Diana and Perri (b) where it is shown that, when $\mu_{2,y} > \sigma_u^2 C_w^{-2}$, with C_w denoting the coefficient of variation of the variable indicated in the subscript, and for ξ minimizing $\mathbb{V}(\hat{\mu}_3)$, the following ordering is valid

$$\mathbb{V}(\hat{\mu}_1) \geq \mathbb{V}(\hat{\mu}_2|\alpha) \geq \mathbb{V}(\hat{\mu}_{EH}) > \mathbb{V}(\hat{\mu}_{PB}) > \mathbb{V}(\hat{\mu}_3) \quad \text{for } 0 \leq \alpha \leq 0.5$$

$$\mathbb{V}(\hat{\mu}_2|\alpha) > \mathbb{V}(\hat{\mu}_1) > \mathbb{V}(\hat{\mu}_{EH}) > \mathbb{V}(\hat{\mu}_{PB}) > \mathbb{V}(\hat{\mu}_3) \quad \text{for } 0.5 < \alpha < 1.$$

For the above models, the Δ measure take the form

$$\Delta_{PB} = \sigma_u^2, \quad \Delta_{EH} = \frac{\mu_{2,y}}{C_w^2}, \quad \Delta_1 = \mu_{2,u} l_w + \mu_{2,y} (l_w - 1) - \mu_u^2$$

$$\Delta_2 = \Delta_1 + \frac{(2\alpha - 1)a + 2\alpha(1 - \alpha)b}{(1 - \alpha)^2} \mu_u, \quad \Delta_3 = \frac{\mu_{2,y} \sigma_w^2 \sigma_u^2}{\mu_w^2 \sigma_u^2 + \mu_{2,y} \sigma_w^2}$$

where $a = \mu_u(l_u l_w - 1)$ and $b = \mu_y(l_w - 1)$ with $l_u = \mu_{2,u}/\mu_u^2$, $l_w = \mu_{2,w}/\mu_w^2$. Comparing the measures we find that the multiplicative model is more protective than the additive one if the condition $\mu_{2,y} > \sigma_u^2 C_w^{-2}$ holds, whereas Δ_1 is always greater than Δ_{EH} since

$$\Delta_1 - \Delta_{EH} = l_u l_w - 1.$$

Similarly, it can be demonstrated that $\Delta_{PB} > \Delta_3$. Therefore, the following ordering holds: $\Delta_1 > \Delta_{EH} > \Delta_{PB} > \Delta_3$. Note that the variances are ordered in the same manner but with an opposite meaning: the more protective the method, the less efficient the estimator. The analysis for the privacy protection induced by the mixed2 model is not so straightforward. It can be proved however that Δ_2 is an increasing function of α in $(0, 1)$ with $\lim_{\alpha \rightarrow 1^-} \Delta_2 = \infty$ and $\Delta_2 = \Delta_{EH}$ for $\alpha = 0$. Then $\Delta_2 > \Delta_{EH} > \Delta_{PB}$ for $\alpha \in (0, 1)$. Moreover, a value of α exists, say $\alpha_0 = (a + b - \sqrt{a^2 + b^2})/2b$ in the interval $(0, 0.5)$, such that $\Delta_2 = \Delta_1$, $\Delta_2 > \Delta_1$ [$\Delta_1 > \Delta_2$] for $\alpha > \alpha_0$ [$\alpha < \alpha_0$]. From efficiency considerations, it follows that for $\alpha \in (\alpha_0, 0.5)$ the mixed2 model outperforms mixed1 model both in efficiency and privacy protection.

The results given in (7) and (8) can be easily extended to the case where $p \in (0, 1)$ and, thus, a direct response on Y is possible. In this situation we find

$$\Delta = (1 - p) \mathbb{E} \left(\frac{S - c_0}{h_0} - Y \right)^2$$

which well emphasizes the (unsurprising) fact that privacy is less protected when a direct response is allowed.

When the auxiliary variable X is used, the variance of the estimators decreases according to expression (3), while the confidentiality of the respondents is expected to decrease as well. This aspect is not captured by the Zaizai et al. (2009) measure. To overcome this drawback, the following measure proposed by Diana and Perri (a) can be employed

$$\tau = 1 - \rho_{y.vz}^2 = 1 - \frac{\rho_{yv}^2 + \rho_{yz}^2 - 2\rho_{yv}\rho_{yz}\rho_{vz}}{1 - \rho_{vz}^2} \quad (9)$$

$\rho_{y.vz}$ being the multiple correlation coefficient. This is a normalized and intuitive measure since, according to the literature on the topic, it indicates how closely the true values of the sensitive variable can be estimated by the perturbed ones. Maximum privacy protection is attained when $\tau = 1$ since $\rho_{y.vz}^2 = 0$. On the contrary, as τ approaches zero privacy protection declines and respondents are expected to become less collaborative. Besides (4) and (5), the determination of τ measure requires additional calculations for the covariance terms, namely

$$\sigma_{yv} = p\sigma_{xy} + (1 - p)\sigma_{yr}, \quad \sigma_{yz} = p\sigma_y^2 + (1 - p)\sigma_{ys}.$$

For illustrative purposes, let us suppose we collect direct information on X and scramble the sensitive response by resorting to the additive and multiplicative models. Then, after some algebra, we get

$$\tau = 1 + \frac{\kappa\rho_{xy}^4 + (\kappa^2 + 2\kappa - 1)\rho_{xy}^2 - \kappa}{1 - \kappa\rho_{xy}^2} \quad (10)$$

where $\kappa = \frac{\sigma_y^2}{\sigma_u^2 + \sigma_y^2}$ for the additive model and $\kappa = \frac{\mu_w^2 \sigma_y^2}{\mu_{2,w} \mu_{2,y} - \mu_w^2 \mu_y^2}$ for the multiplicative one.

When no auxiliary information is used, from (9) we get $\tau = 1 - \rho_{yz}^2$ which reduces to

$$\tau = 1 - h^2 \frac{\sigma_y^2}{\sigma_z^2}$$

if the variable Y is linearly perturbed whatever $p \in (0, 1)$. The models shown meet this requirement. Therefore, from (3), the more efficient the estimates, the less protective the coding procedure. Note that the expression for κ in (10) corresponds to $h^2 \sigma_y^2 / \sigma_z^2$.

Finally, when a direct response on Y is allowed with probability $p \in (0, 1)$, the privacy measured by τ can be better guaranteed through a suitable choice of p . For instance, the response model

$$Z = \begin{cases} Y, & \text{with probability } p \\ S = Y + U, & \text{with probability } 1 - p \end{cases}$$

is more protective (in the τ sense) than the simple model $Z = Y + U$ if $p < 2 - l_u$. If the scrambling variable U is chosen such that $l_u > 2$, then the condition is never met and respondents confidentiality can be increased by avoiding direct questioning.

5 Discussion

The idea of releasing algebraically perturbed data rather than the exact values of a quantitative variable can be particularly useful to combat deliberate misreporting or untruthful responses when people are surveyed on sensitive issues. The paper shows how to treat scrambled response models in a general framework in order to estimate the mean of a sensitive variable when an auxiliary variable can be used directly at the estimation stage. The superiority of estimators based on auxiliary information is clearly shown, whatever the adopted scrambling procedure.

A relevant aspect when dealing with coding methods is the degree of privacy protection assured to the respondents. Methods which produce high efficiency of the estimates tend to be more intrusive and jeopardize privacy to a much greater extent. The key issue when choosing a model is to find a right trade-off between privacy protection and efficiency in the estimates, hence the importance of measuring the degree of privacy protection correctly. Different measures are discussed in the literature two of which are mentioned in the paper. The main concern is that the measures can yield different conclusions and do not always allow us to capture certain relevant aspects such as the effect of the auxiliary information. Moreover, even if they can provide the researcher with an indication of privacy disclosure it is questionable whether the respondents themselves are able to measure the extent of privacy disclosure offered by a scrambling design. Whatever the privacy measure, respondents usually have a subjective impression regarding disclosure which can be

higher or lower than that measured by an index. In this sense, an empirical study would certainly be of interest to ascertain and quantify the difference between the perceived and measured degree of respondents' privacy protection.

Finally, we observe that since respondents can not fully appreciate the level of privacy protection assured by a scrambled response method, researchers may be more inclined towards designs allowing more efficient estimates rather than ones that offer high protection in theory. On the other hand, simple coding mechanisms should be preferred since understanding how to perturb the responses algebraically may be not easy. This may be a limitation for the application of the scrambled methods since their correct execution can only be performed by highly educated people unless a computer support is provided.

Acknowledgements Work supported by the Italian Ministry of University and Research, MIUR-PRIN 2007: "Efficient use of auxiliary information at the design and at the estimation stage of complex surveys: methodological aspects and applications for producing official statistics".

References

1. Arnab, R., Singh S.: Randomized response techniques: an application to the Botswana AIDS impact survey. *J. Stat. Plann. Inference* **140**, 941–953 (2010)
2. Bar-Lev, S.K., Bobovitch, E., Boukai, B.: A note on randomized response models for quantitative data. *Metrika* **60**, 255–260 (2004)
3. Cochran, W.G.: *Sampling Techniques*. John Wiley & Sons, New York (1977)
4. Diana, G., Perri, P.F.: Estimation of finite population mean using multi-auxiliary information. *Metron* **LXV** 99–112 (2007)
5. Diana, G., Perri, P.F.: Estimating a sensitive proportion through randomized response procedures based on auxiliary information. *Stat. Pap.* **50**, 661–672 (2009)
6. Diana, G., Perri, P.F.: A class of estimators for quantitative data. *Stat. Pap.* (a, in press)
7. Diana, G., Perri, P.F.: New scrambled response models for estimating the mean of a sensitive quantitative character. *J. Appl. Stat.* (b, in press)
8. Eichhorn, B., Hayre, L.S.: Scrambled randomized response methods for obtaining sensitive quantitative data. *J. Stat. Plann. Inference* **7**, 307–316 (1983)
9. Gjestavang, C.R., Singh, S.: An improved randomized response model: estimation of mean. *J. Appl. Stat.* **36**, 1361–1367 (2009)
10. Gupta, S., Gupta, B., Singh, S.: Estimation of sensitive level of personal interview survey questions. *J. Stat. Plann. Inference* **100**, 239–247 (2002)
11. Krumpal, I.: (2010): Estimating the prevalence of xenophobia and anti-semitism in Germany: a comparison of the randomized response technique and direct questioning in the telephone survey mode. Unpublished Manuscript, Institute of Sociology, University of Leipzig (2010)
12. Lara, D., Strickler, J., Olavarrieta, C.D., García, S.G., Ellertson, C.: Measuring induced abortion in Mexico. A comparison of four methodologies. *Soc. Meth. R.* **32**, 529–558 (2004)
13. Lara, D., García, S.G., Ellertson, C., Camlin, C., Suárez, J.: The measure of induced abortion in Mexico using random response technique. *Soc. Meth. R.* **35**, 279–301 (2006)
14. Lensvelt-Mulders, G.J.L.M., Hox, J.J., van der Heijden, P.G.M., Maas, C.J.M.: Meta-analysis of randomized response research: thirty-five years of validation. *Soc. Meth. R.* **33**, 319–348 (2005)
15. Lensvelt-Mulders, G.J.L.M., van der Heijden, P.G.M., Laudy, O., van Gils, G.: A validation of a computer-assisted randomized response survey to estimate the prevalence of fraud in social security. *J. R. Stat. Soc., Ser. A* **169**, 305–318 (2006)

16. Ostapczuk, M., Musch, J., Mashagen, M.: A randomized-response investigation of the education effect in attitudes towards foreigners. *Eur. J. Soc. Psychol.* **39**, 920–931 (2009)
17. Pollock, K.H., Bek, Y.: A comparison of three randomized response models for quantitative data. *J. Am. Stat. Assoc.* **71**, 884–886 (1976)
18. Saha, A.: A simple randomized response technique in complex surveys. *Metron* **LXV**, 59–66 (2007)
19. van der Heijden, P.G.M., van Gils, G., Bouts, J., Hox, J.J.: A comparison of randomized response, computer-assisted self-interview, and face-to-face direct questioning. *Soc. Meth. R.* **28**, 505–537 (2000)
20. van der Heijden, P.G.M., Böckenholt, U.: Applications of randomized response methodology in e-commerce. In: Jank, W., Shmueli, G. (eds.) *Statistical Methods in e-Commerce Research*, pp. 401–416. Wiley, New York (2008)
21. Warner, S.L.: Randomized response: a survey technique for eliminating evasive answer bias. *J Am Stat Assoc* **60**, 63–69 (1965)
22. Willenborg, T., de Waal, T.: *Statistical Disclosure Control in Practice*. Springer, Berlin (1996)
23. Zaizai, Y., Jingyu, W., Junfeng, L.: An efficiency and protection degree-based comparison among the quantitative randomized response strategies. *Commun. Stat., Simulation Comput.* **38**, 400–408 (2009)