

# Use of Randomized Response Techniques When Data are Obtained from Two Frames

Maria del Mar Rueda\*, Antonio Arcos and Beatriz Cobo

Department of Statistics and Operational Research, University of Granada, Spain

Received: 12 Jul. 2014, Revised: 13 Oct. 2014, Accepted: 14 Oct. 2014

Published online: 1 Apr. 2015

**Abstract:** The methodology of randomized response has advanced considerably in recent years. Nevertheless, to date all the proposed estimators with randomized response techniques have been based on the hypothesis of the availability of a unique and complete list of units forming the target population to be used as a sampling frame. In this paper, we present a new procedure aimed at determining a population total using a model of randomized response when data are obtained from two frames. We introduce different ways of combining estimates obtained from the different frames and propose unbiased estimators, with an analytic expression for their variances. Estimates for the variances are also obtained, applying analytical formulas such as those based on resampling technologies. A simulation study illustrates the behaviour of the estimator using diverse randomization devices.

**Keywords:** Randomized response, sampling design, scrambling distribution, surveys

## 1 Introduction

In psychological and social surveys, people often do not respond truthfully when asked personal or sensitive questions. Obtaining valid and reliable information depends on the cooperation of the respondents, and this depends on the confidentiality of their responses. Any research study that uses self-report measures runs the risk of response bias. There is ample empirical evidence that respondents systematically over-report socially desirable behaviour and attitudes and systematically under-report socially undesirable ones. ([1]).

[2] developed a data collection procedure, the randomized response (RR) technique, which allows researchers to obtain sensitive information while guaranteeing privacy to respondents. This method encourages greater cooperation from respondents and reduces their motivation to falsely report their attitudes. The most important claim made for RR is that it yields more valid point estimates of sensitive behaviour: there have been many reports that RR achieves more accurate estimates of the prevalence of socially undesirable behaviour than when sensitive questions are asked directly ([3]). However, using RR incurs extra costs (RR techniques produce larger sampling variances, which leads to reduced power and thus necessitates larger

samples) and RR questions present increased complexity compared to more conventional forms of data collection. The advantage of using RR, i.e., the greater accuracy of the population estimates obtained, will only outweigh these extra costs if the estimates are substantially better than those derived from straightforward question-and-answer designs ([4]).

Warner's study generated a rapidly-expanding body of research literature on alternative techniques for eliciting suitable RR schemes in order to estimate a population proportion (see [5], [6], ...). [7] presents a good review of pioneering work in the field of RR.

All the estimators currently used in RR are based on the hypothesis of the availability of a unique and complete list of units forming the target population to be used as a sampling frame. In many situations, however, there is no single frame that covers all the population; on the other hand, there are several sampling frames whose joint extension covers the population of interest. In this situation, we can create a new frame, combining those available and deleting the intersections between them. Nevertheless, it may be more practical to take samples from the different sampling frames and then combine the information from the samples to estimate population quantities. For example, the National Incidence Study of Child Abuse and Neglect is a national survey to estimate

\* Corresponding author e-mail: [mrueda@ugr.es](mailto:mrueda@ugr.es)

the number and the characteristics of maltreated children in the United States. This study uses a multiple frame design to broaden the coverage of reporting sources for maltreated children. In this design, the first frame is a list frame of maltreated children investigated by child protection agencies. However, the coverage of these agencies is incomplete because some maltreated children may not be investigated by them, and so a second frame is employed, including the children observed by non-official agencies and classified as possibly maltreated (see [8]).

Multiple frame surveys are obviously useful when no single frame covers the whole target population but the union of several available frames does. They also have other advantages. In fact, [9] introduced dual-frame surveys as a cost-saving device, showing that they can often achieve the same precision as a single-frame survey at a greatly reduced cost.

One of the main advantages of using multiple frames is that different information withdrawal procedures can be used; a response protection procedure may be possible in one frame but not in another. This would allow us to combine the advantages of RR surveys with those obtained from direct answer surveys. In this respect, [10] consider a dual sampling scheme with direct questioning and use RR to investigate violations of social security regulations.

This paper introduces the theory of the random response technique in the presence of multiple frames. We consider two different randomization models and propose several unbiased estimators to obtain the total of a sensitive quantitative variable.

## 2 RR Techniques

In this section, we present some well-known RR techniques relevant to this discussion. Consider a finite population  $U = \{1, \dots, i, \dots, N\}$ , consisting of  $N$  different elements. Let  $y_i$  be the value of the sensitive aspect under study for the  $k$ th population element. Our aim is to estimate the finite population total  $Y = \sum_{i=1}^N y_i$  of the variable of interest  $y$  or the population mean  $\bar{Y} = \frac{1}{N} \sum_{i=1}^N y_i$ . If we can estimate the proportion of the population presenting a certain stigmatised behaviour  $C$ , the variable  $y_i$  takes the value 1 if  $i \in G_C$  (the group with the stigmatised behaviour) and the value zero otherwise.

[2] developed the first RR data collection procedure, as follows: from the population  $U$ , a sample of  $n$  units is selected by the method of simple random sampling with replacement (SRSWR). Each of the selected units performs a RR trial as follows. The respondent is instructed to select a card at random from a pack of cards consisting of two types of cards with known proportions. Card type 1, with proportion  $\theta$  marked I, belongs to group A and card type 2, with proportion  $1 - \theta$  marked II, does not belong to group A. The respondent must truthfully answer yes or no. The experiment is performed in the absence of the interviewer and hence confidentiality

is maintained because the interviewer will not know which question the respondent has answered. Let  $Z_i$  denote the scrambled response from the  $i$ -th respondent. This variable takes the value 1(0) if the response yes (no) was obtained. Denoting  $E_R$  ( $V_R$ ) as the operator of expectation (variance) for the RR, we have  $E_R(Z_i) = y_i\theta + (1 - y_i)(1 - \theta)$  and  $V_R(Z_i) = \theta(1 - \theta)$ .

The revised randomized response  $R_i = \frac{z_i - 1 - \theta}{2\theta - 1}$  is an unbiased estimator of  $y_i$  and satisfies

$$E(\bar{R}) = E\left(\frac{1}{n} \sum_{i=1}^n R_i\right) = \bar{Y}$$

under simple random sampling, that is, the sample mean of the revised RR values is an unbiased estimator of the proportion of the population presenting the stigmatised behaviour.

[11] proposed a modification of the Warner model. In this case, the RR trial consists of two questions: one related to the sensitive character ( $y$ ) and the other to a neutral character ( $q$ ) such that: (i) I possess character C and (ii) I possess character Q. The respondent must select either question (i) with probability  $\theta$  or question (ii) with probability  $1 - \theta$ , using a suitable randomization device, and provide the answer Yes or No to the interviewer. This model is also known as the U model [12]. Let us define the RR obtained from the  $i$ -th respondent as  $z_i$ . In this case, the revised RR is

$$R_i = \frac{Z_i - \pi_Q(1 - \theta)}{\theta}$$

where  $\pi_Q$  is the proportion of persons who possess the non-sensitive character in the population, which is assumed to be known.

[13] proposed the H model, which provides greater protection of the interviewees anonymity, without using a complementary question. Each element of the sample is instructed to respond randomly to one of three propositions: (1) the sensitive question; (2) an instruction to say yes; and (3) an instruction to say no. These are chosen with probabilities of  $p_1$ ,  $p_2$  and  $p_3$ , with  $p_1 + p_2 + p_3 = 1$ . In the M model [14], the random mechanism provides  $n$  independent responses with two random components. The D model [15] is analogous to U, with one basic difference: the fact of belonging to the innocuous group is established with a probability of one.

Warner's study generated a rapidly-expanding body of research literature on alternative techniques for eliciting suitable RR schemes in order to estimate a population proportion (see [16], [17], [5], [18], [10], [19], [6], ...).

Standard RR methods are used primarily in surveys which require a binary response to a sensitive question, and seek to estimate the proportion of the population presenting a given (sensitive) characteristic. Nevertheless, some studies have addressed situations in which the response to a sensitive question results in a quantitative variable. [20] extended RR to this case, rather than a

simple Yes or No. In this study, the respondent was asked to select, by means of a randomization device, one of two questions; the sensitive one or an unrelated question, the answers to which were of about the same order of magnitude. In addition, other important randomization devices have been proposed:

1. The Eichhorn and Hayre method [21]. Each respondent selected from a simple random sample is asked to report the scrambled response  $Z_i = Sy_i$  where  $S$  is a scramble variable whose distribution is assumed to be known.
2. The Bar-Levy, Bobovitch and Boukai method [22]. Each respondent from a simple random sample is requested to rotate a spinner unobserved by the interviewer, and if the spinner stops in the shaded area, then the respondent is asked to report the real response on the sensitive variable,  $y_i$ . Otherwise, the respondent is asked to report the scrambled response  $Z_i = Sy_i$ .
3. The FQRR method [23]. Each respondent selected from a simple random sample is provided with a randomization device bearing three types of statements:
  - report the true value of the sensitive variable  $y_i$
  - report the scrambled response  $Z_i + y_iS$
  - report a fixed value  $F$
 with proportions  $p_1, p_2$  and  $p_3$  and where  $S$  is the scramble variable.

Other important RR methods for quantitative variables are given in [24], [25], [26] and [27]. [28], [29] and [30] propose unified approaches to the discrete and continuous models considered.

### 3 Randomized response techniques with data obtained from dual frames

Following [9] suppose that we have two frames  $A$  and  $B$ , which together cover the population  $U$ . Let  $\mathcal{A}$  be the set of population units in frame  $A$  and  $\mathcal{B}$  the set of population units in frame  $B$ . The population of interest,  $\mathcal{U}$ , may be divided into three mutually exclusive domains,  $a = \mathcal{A} \cap \mathcal{B}^c, b = \mathcal{A}^c \cap \mathcal{B}$  and  $ab = \mathcal{A} \cap \mathcal{B}$ . The population units in the overlap domain  $ab$  can be sampled in either survey or in both surveys. Let  $N, N_A, N_B, N_a, N_b, N_{ab}$  be the number of population units in  $\mathcal{U}, A, B, a, b, ab$  respectively.

Let  $\delta_i(a) = 1$  if  $i \in a$  and 0 otherwise,  $\delta_i(b) = 1$  if  $i \in b$  and 0 otherwise and  $\delta_i(ab) = 1$  if  $i \in ab$  and 0 otherwise. Two probability samples  $s_A$  and  $s_B$  of sizes  $n_A$  and  $n_B$ , are drawn independently from frame  $A$  and frame  $B$  under sampling designs  $d_A = (S_A, p_{dA})$  and  $d_B = (S_B, p_{dB})$  respectively. Each design induces first-order inclusion probabilities  $\pi_k^A$  and  $\pi_k^B$ , respectively. The final sample  $s$  is obtained as  $s_A \cup s_B$ .

Let  $y$  be a variable of interest in the population and  $y_k$  its value on unit  $k$ , for  $k = 1, \dots, N$ . Our aim is to estimate

the finite population total  $Y = \sum_{k=1}^N y_k$  of  $y$ , which can be written as

$$Y = Y_a + Y_{ab} + Y_b. \tag{1}$$

We will assume that in each frame it is possible to use a different randomized response procedure.

In order to consider a wide variety of RR procedures, we consider the unified approach given by [28]. The interviews of individuals in the sample  $s_A$  are conducted in accordance with the RR model used in this frame (denoted by  $RRA$ ). For each  $i \in s_A$  the  $RRA$  induces a random variable  $Z_{Ai}$  so that the revised randomized response  $R_{Ai}$  is an unbiased estimation of  $y_i$ , the real value of the sensitive quantitative variable. Similarly, each respondent in the sample  $s_B$  is requested to report the scrambled response  $Z_{Bi}$  with the revised randomized response  $R_{Bi}$ . The RR model used in frame  $B$  is noted as  $RRB$ . We consider  $RRA$  and  $RRB$  to be independent randomized devices such that the respective revised randomized responses  $R_{Ai}$  and  $R_{Bi}$  satisfy the conditions (see [5]):

$$\begin{aligned} E_R(R_{Ai}) &= y_i, V_R(R_{Ai}) = \sigma_{Ai}^2, C_R(R_{Ai}, R_{Aj}) = 0, \\ E_R(R_{Bi}) &= y_i, V_R(R_{Bi}) = \sigma_{Bi}^2, C_R(R_{Bi}, R_{Bj}) = 0. \end{aligned}$$

Most RR models for qualitative or quantitative characteristics satisfy these conditions.

Since  $y_i$ 's are not known for every  $i \in s$  we propose two unbiased estimators for the total.

### 4 Single frame estimators

Classical single frame (SF) methods ([31]) estimate the population total by treating all observations as though they had been sampled from a single frame, and the sampling weights of observation in the intersection domain are modified according to their inclusion probability in each sample.

The unit  $i$  in the intersection domain  $ab$  could be selected both in the samples from frame  $A$  and in those from frame  $B$ , and so the expected number of times it can be selected is  $\pi_i^A + \pi_i^B$ . Thus, if  $w_{Ai} = \frac{1}{\pi_i^A}$  and  $w_{Bi} = \frac{1}{\pi_i^B}$ , the adjusted weights for the units sampled in frame  $A$  are:

$$\tilde{w}_{SF_i} = \begin{cases} w_{Ai} & \text{if } i \in a \\ w_{Bi} & \text{if } i \in b \\ (1/w_{Ai} + 1/w_{Bi})^{-1} & \text{if } i \in ab \end{cases}$$

Using this idea, we propose a new methodology to apply RR techniques.

*Theorem 1. An unbiased estimator of the proposed total  $Y$  is given by*

$$e_{SF}(Y) = \sum_{i \in s_A} \tilde{w}_{SF_i} R_{Ai} + \sum_{i \in s_B} \tilde{w}_{SF_i} R_{Bi} \tag{2}$$

Proof.

Writing  $E_d, V_d$  as the expectation-variance operators for any sampling design  $d$  and  $E_R, V_R$  as the expectation-variance operators over the RR device, we have

$$\begin{aligned}
 E(e_{SF}(Y)) &= E_d E_R \left( \sum_{i \in s_A} \tilde{w}_{SF_i} R_{Ai} + \sum_{i \in s_B} \tilde{w}_{SF_i} R_{Bi} \right) \quad (3) \\
 &= E_d \left( \sum_{i \in s_A} \tilde{w}_{SF_i} E_R(R_{Ai}) \right) + E_d \left( \sum_{i \in s_B} \tilde{w}_{SF_i} E_R(R_{Bi}) \right) \\
 &= E_d \left( \sum_{i \in s_A} \tilde{w}_{SF_i} y_i \right) + E_d \left( \sum_{i \in s_B} \tilde{w}_{SF_i} y_i \right) \\
 &= \sum_{i \in U_A} \tilde{w}_{SF_i} y_i E_d(I_i(A)) + \sum_{i \in U_B} \tilde{w}_{SF_i} y_i E_d(I_i(B))
 \end{aligned}$$

where  $\forall i \in U, I_i(A) = 1$  if  $i \in s_A$  and 0 otherwise,  $I_i(B) = 1$  if  $i \in s_B$  and 0 otherwise,  $E_d(I_i(A)) = \pi_i^A$  and  $E_d(I_i(B)) = \pi_i^B$ .

Thus

$$\begin{aligned}
 E(e_{SF}(Y)) &= \sum_{i \in U_a} \frac{y_i}{\pi_i^A} \pi_i^A + \sum_{i \in U_{ab}} \frac{y_i}{\pi_i^A + \pi_i^B} \pi_i^A \quad (4) \\
 &+ \sum_{i \in U_b} \frac{y_i}{\pi_i^B} \pi_i^B + \sum_{i \in U_{ab}} \frac{y_i}{\pi_i^A + \pi_i^B} \pi_i^B \\
 &= Y_a + Y_b + Y_{ab} = Y
 \end{aligned}$$

Then,  $e_{SF}(Y)$  is an unbiased estimator of the population total  $Y$ .

*Theorem 2. The variance of  $e_{SF}(Y)$  is given by*

$$\begin{aligned}
 V(e_{SF}(Y)) &= \sum_{i \in U_A} \sigma_{Ai}^2 \tilde{w}_{SF_i}^2 \pi_i^A + \sum_{i \in U_B} \sigma_{Bi}^2 \tilde{w}_{SF_i}^2 \pi_i^B \quad (5) \\
 &+ \sum_{i,j \in U_A} y_i y_j (\tilde{w}_{SF_i} \tilde{w}_{SF_j} \pi_{ij}^A - 1) \\
 &+ \sum_{i,j \in U_B} y_i y_j (\tilde{w}_{SF_i} \tilde{w}_{SF_j} \pi_{ij}^B - 1)
 \end{aligned}$$

where  $\pi_{ij}^A$  and  $\pi_{ij}^B$  denote the second-order inclusion probabilities in each frame.

Proof.

$$\begin{aligned}
 V(e_{SF}(Y)) &= E_d V_R(e_{SF}(Y)) + V_d E_R(e_{SF}(Y)) \\
 &= E_d \left( \sum_{i \in s_A} \tilde{w}_{SF_i}^2 V_R(R_{Ai}) \right) + E_d \left( \sum_{i \in s_B} \tilde{w}_{SF_i}^2 V_R(R_{Bi}) \right) \\
 &+ V_d \left( \sum_{i \in s_A} \tilde{w}_{SF_i} E_R(R_{Ai}) \right) + V_d \left( \sum_{i \in s_B} \tilde{w}_{SF_i} E_R(R_{Bi}) \right) \\
 &= \sum_{i \in U_A} \tilde{w}_{SF_i}^2 \sigma_{Ai}^2 E_d(I_i(A)) + \sum_{i \in U_B} \tilde{w}_{SF_i}^2 \sigma_{Bi}^2 E_d(I_i(B)) \\
 &+ \sum_{i,j \in U_A} y_i y_j \left( \sum_{s_A \ni i,j} \tilde{w}_{SF_i} \tilde{w}_{SF_j} p_{dA}(s_A) - 1 \right) \\
 &+ \sum_{i,j \in U_B} y_i y_j \left( \sum_{s_B \ni i,j} \tilde{w}_{SF_i} \tilde{w}_{SF_j} p_{dB}(s_B) - 1 \right) \\
 &= \sum_{i \in U_A} \tilde{w}_{SF_i}^2 \sigma_{Ai}^2 \pi_i^A + \sum_{i \in U_B} \tilde{w}_{SF_i}^2 \sigma_{Bi}^2 \pi_i^B \\
 &+ \sum_{i,j \in U_A} y_i y_j (\tilde{w}_{SF_i} \tilde{w}_{SF_j} \pi_{ij}^A - 1) \\
 &= \sum_{i,j \in U_B} y_i y_j (\tilde{w}_{SF_i} \tilde{w}_{SF_j} \pi_{ij}^B - 1)
 \end{aligned}$$

The variance of the estimator is composed of four terms, the last two of which depend on the sampling designs  $d_A$  and  $d_B$  and the  $y_i$  values in each frame. These terms are common to all of the RR models. The first and second terms depend on the sampling design and also on the random mechanism used in each frame.

### 5 Estimating the variance of the proposed estimator

From expression (5), and using tools derived from sampling theory, we can obtain an unbiased estimator for  $V(e_{SF}(Y))$ . The procedure for this depends on the sample design and the randomization method used, and will be different in each situation. The following procedure is applied for some specific RR models.

#### 5.1 Qualitative methods

An analytical expression for the variance estimator can be obtained straightforwardly to estimate the proportion of individuals with a given feature.

*Theorem 3. If the variable of interest  $y$  is dichotomous, an unbiased estimator of the variance of  $e_{SF}(Y)$  is given by*

$$\begin{aligned}
 \hat{V}(e_{SF}(Y)) &= \sum_{i \in s_A} R_i(R_i - 1) \tilde{w}_{SF_i}^2 + \sum_{i \in s_B} R_i(R_i - 1) \tilde{w}_{SF_i}^2 \quad (6) \\
 &+ \sum_{i,j \in s_A} R_i R_j \frac{\tilde{w}_{SF_i} \tilde{w}_{SF_j} \pi_{ij}^A - 1}{\pi_{ij}^A} \\
 &+ \sum_{i,j \in s_B} R_i R_j \frac{\tilde{w}_{SF_i} \tilde{w}_{SF_j} \pi_{ij}^B - 1}{\pi_{ij}^B}
 \end{aligned}$$

Proof.

$$\begin{aligned}
 E(\widehat{V}(e_{SF}(Y))) &= E_d E_R \left( \sum_{i \in s_A} R_i(R_i - 1) \tilde{w}_{SF_i}^2 \right) \\
 &+ \sum_{i \in s_B} R_i(R_i - 1) \tilde{w}_{SF_i}^2 \\
 &+ E_d E_R \left( \sum_{i,j \in s_A} R_i R_j \frac{\tilde{w}_{SF_i} \tilde{w}_{SF_j} \pi_{ij}^A - 1}{\pi_{ij}^A} \right) \\
 &+ \sum_{i,j \in s_B} R_i R_j \frac{\tilde{w}_{SF_i} \tilde{w}_{SF_j} \pi_{ij}^B - 1}{\pi_{ij}^B}
 \end{aligned} \tag{7}$$

We consider the two terms separately.

For a qualitative characteristic an unbiased estimator of  $V_R(R_{Ai})$  is  $R_i(R_i - 1)$  because  $E_R(R_i(R_i - 1)) = E_R(R_i^2) - E_R(R_i) = V_R(R_{Ai}) + y_i^2 - y_i = V_R(R_{Ai})$  as  $y_i^2 = y_i$ . Thus

$$\begin{aligned}
 E_d E_R \left( \sum_{i \in s_A} R_i(R_i - 1) \tilde{w}_{SF_i}^2 + \sum_{i \in s_B} R_i(R_i - 1) \tilde{w}_{SF_i}^2 \right) &= \\
 = E_d \left( \sum_{i \in s_A} \sigma_{Ai}^2 \tilde{w}_{SF_i}^2 \right) + E_d \left( \sum_{i \in s_B} \sigma_{Bi}^2 \tilde{w}_{SF_i}^2 \right) &= \\
 = \sum_{i \in U_A} \sigma_{Ai}^2 \tilde{w}_{SF_i}^2 \pi_i^A + \sum_{i \in U_B} \sigma_{Bi}^2 \tilde{w}_{SF_i}^2 \pi_i^B.
 \end{aligned}$$

On the other hand  $E_R(R_i R_j) = Cov(R_i, R_j) + E(R_i)E(R_j) = y_i y_j$  because  $Cov(R_i, R_j) = 0$ . Thus

$$\begin{aligned}
 E_d E_R \left( \sum_{i,j \in s_A} R_i R_j \frac{\tilde{w}_{SF_i} \tilde{w}_{SF_j} \pi_{ij}^A - 1}{\pi_{ij}^A} + \right. \\
 \left. \sum_{i,j \in s_B} R_i R_j \frac{\tilde{w}_{SF_i} \tilde{w}_{SF_j} \pi_{ij}^B - 1}{\pi_{ij}^B} \right) &= \\
 E_d \left( \sum_{i,j \in s_A} y_i y_j \frac{\tilde{w}_{SF_i} \tilde{w}_{SF_j} \pi_{ij}^A - 1}{\pi_{ij}^A} + \right. \\
 \left. \sum_{i,j \in s_B} y_i y_j \frac{\tilde{w}_{SF_i} \tilde{w}_{SF_j} \pi_{ij}^B - 1}{\pi_{ij}^B} \right) &= \\
 \sum_{i,j \in U_A} y_i y_j (\tilde{w}_{SF_i} \tilde{w}_{SF_j} \pi_{ij}^A - 1) + \sum_{i,j \in U_B} y_i y_j (\tilde{w}_{SF_i} \tilde{w}_{SF_j} \pi_{ij}^B - 1)
 \end{aligned}$$

### 5.2 Quantitative methods

An expression for the unbiased estimator for the common variance that is valid for any mechanism of randomization cannot be obtained. Here, for illustrative purposes, we present an estimator for the variance for a particular model of randomization, the EH model in each frame.

Each respondent in sample  $s_A$  is asked to report the scrambled response  $Z_{Ai} = S_A y_i$  where  $y_i$  is the real value

of the sensitive quantitative variable, and  $S_A$  is the scrambling variable. Similarly, each respondent in the sample  $s_B$  is asked to report the scrambled response  $Z_{Bi} = S_B y_i$  where  $S_B$  is another scrambling variable.  $S_A$  and  $S_B$  are assumed to have different known distributions, that is,  $E(S_A) = \mu_A$ ,  $V(S_A) = \sigma_A^2$ ,  $E(S_B) = \mu_B$  and  $V(S_B) = \sigma_B^2$  are assumed to be known and positive.

*Theorem 4.* Under the above randomized device in each frame, an unbiased estimator of the variance of  $e_{SF}(Y)$  is given by

$$\begin{aligned}
 \widehat{V}(e_{SF}(Y)) &= \\
 \frac{CV(S_A)^2}{1 + CV(S_A)^2} \sum_{i \in s_A} R_{Ai}^2 \tilde{w}_{SF_i}^2 + \frac{CV(S_B)^2}{1 + CV(S_B)^2} \sum_{i \in s_B} R_{Bi}^2 \tilde{w}_{SF_i}^2 &+ \\
 \sum_{i,j \in s_A} R_i R_j \frac{\tilde{w}_{SF_i} \tilde{w}_{SF_j} \pi_{ij}^A - 1}{\pi_{ij}^A} + \sum_{i,j \in s_B} R_i R_j \frac{\tilde{w}_{SF_i} \tilde{w}_{SF_j} \pi_{ij}^B - 1}{\pi_{ij}^B} &
 \end{aligned} \tag{8}$$

where  $CV(S_A)$  and  $CV(S_B)$  are the variation coefficient of scrambled variables  $S_A$  and  $S_B$  respectively.

Proof.

For these devices, we have:  $R_i = \frac{1}{\mu_A} Z_{Ai}$  for  $i \in s_A$  and  $R_i = \frac{1}{\mu_B} Z_{Bi}$  for  $i \in s_B$ ,  $V_R(R_i) = y_i^2 CV(S_A)^2$  for  $i \in s_A$  and  $V_R(R_i) = y_i^2 CV(S_B)^2$  for  $i \in s_B$ .

$$\begin{aligned}
 E \left( \frac{CV(S_A)^2}{1 + CV(S_A)^2} \sum_{i \in s_A} R_{Ai}^2 \tilde{w}_{SF_i}^2 + \frac{CV(S_B)^2}{1 + CV(S_B)^2} \sum_{i \in s_B} R_{Bi}^2 \tilde{w}_{SF_i}^2 \right) &= \\
 = \frac{CV(S_A)^2}{1 + CV(S_A)^2} E_d \left( \sum_{i \in s_A} E_R(R_{Ai}^2) \tilde{w}_{SF_i}^2 \right) &+ \\
 + \frac{CV(S_B)^2}{1 + CV(S_B)^2} E_d \left( \sum_{i \in s_B} E_R(R_{Bi}^2) \tilde{w}_{SF_i}^2 \right) &= \\
 = \frac{CV(S_A)^2}{1 + CV(S_A)^2} \sum_{i \in U_A} y_i^2 (CV(S_A)^2 + 1) \tilde{w}_{SF_i}^2 \pi_i^A &+ \\
 + \frac{CV(S_B)^2}{1 + CV(S_B)^2} \sum_{i \in U_B} y_i^2 (CV(S_B)^2 + 1) \tilde{w}_{SF_i}^2 \pi_i^B &= \\
 = \sum_{i \in U_A} \sigma_{Ai}^2 \tilde{w}_{SF_i}^2 \pi_i^A + \sum_{i \in U_B} \sigma_{Bi}^2 \tilde{w}_{SF_i}^2 \pi_i^B.
 \end{aligned}$$

On the other hand:

$$\begin{aligned}
 E_d E_R \left( \sum_{i,j \in s_A} R_i R_j \frac{\tilde{w}_{SF_i} \tilde{w}_{SF_j} \pi_{ij}^A - 1}{\pi_{ij}^A} + \right. \\
 \left. \sum_{i,j \in s_B} R_i R_j \frac{\tilde{w}_{SF_i} \tilde{w}_{SF_j} \pi_{ij}^B - 1}{\pi_{ij}^B} \right) &= \\
 \sum_{i,j \in U_A} E_R(R_i R_j) (\tilde{w}_{SF_i} \tilde{w}_{SF_j} \pi_{ij}^A - 1) +
 \end{aligned}$$

$$\sum_{i,j \in U_B} E_R(R_i R_j) (\tilde{w}_{SF_i} \tilde{w}_{SF_j} \pi_{ij}^B - 1)$$

and from  $E_R(R_i R_j) = cov_R(R_i, R_j) + E_R(R_i)E_R(R_j) = y_i y_j$  we obtain the required result.

### 5.3 Resampling methods for estimating the variance

To calculate the above estimators, we need to know the second-order inclusion probabilities of each pair of units in samples  $s_A$  and  $s_B$  and the unbiased estimators for  $\sigma_{A_i}^2$  and  $\sigma_{B_i}^2$ . In some complex sampling designs and RR techniques it is difficult to obtain these values (the unbiased estimator for  $\sigma_i^2$  for some continuous techniques can be seen in [5]. Another, simpler alternative is to use resampling techniques such as jackknife, half-samples or bootstrap (see [32]).

Because samples A and B are independent, the variance is

$$\begin{aligned} V(e_{SF}(Y)) &= V\left(\sum_{i \in s_A} \tilde{w}_{SF_i} \frac{Z_{Ai}}{\mu_A}\right) + V\left(\sum_{i \in s_B} \tilde{w}_{SF_i} \frac{Z_{Bi}}{\mu_B}\right) \\ &= V(e_{A1}) + V(e_{B1}) \end{aligned}$$

and we can estimate the variances separately by means of each of these techniques.

For example, we can consider the jackknife estimator

$$\begin{aligned} v_J(e_{SF}(Y)) &= \frac{n_A - 1}{n_A} \sum_{i=1}^{n_A} (e_{A1}(i) - e_{A1}(J))^2 \\ &+ \frac{n_B - 1}{n_B} \sum_{i=1}^{n_B} (e_{B1}(i) - e_{B1}(J))^2 \end{aligned} \tag{9}$$

where  $e_{A1}(i)$  is the estimator  $e_{A1}$  after dropping the unit  $i$  from the given sample  $s_A$ ,  $e_{A1}(J)$  is the sample mean of the values  $e_{A1}(i)$ ;  $e_{B1}(i)$  and  $e_{B1}(J)$  are defined similarly.

## 6 Averaging the estimates from the overlapping domain

In this section we consider an alternative approach to obtain estimators by combining the randomized values in each frame.

For a general survey with two overlapping frames, the population total can be written as the sum of the total populations of three domains:

$$Y = Y_a + Y_{ab} + Y_b = Y_a + \eta Y_{ab} + (1 - \eta) Y_{ba} + Y_b,$$

where  $0 \leq \eta \leq 1$  is a fixed constant.

A simple way to estimate the population total is to average the domain estimators for domains that are sampled in more than one frame. Several estimators were

defined using this technique (termed the dual frame approach), which was introduced by ([33]).

Using this idea, we propose the estimator

$$\begin{aligned} e_H(Y) &= \sum_{i \in s_A} w_{A_i} \delta_i(a) R_{A_i} + \eta \sum_{i \in s_A} w_{A_i} \delta_i(ab) R_{A_i} \\ &+ (1 - \eta) \sum_{i \in s_B} w_{B_i} \delta_i(ab) R_{B_i} + \sum_{i \in s_B} w_{B_i} \delta_i(b) R_{B_i}. \end{aligned} \tag{10}$$

*Theorem 5.*  $e_H(Y)$  is an unbiased estimator of the total  $Y$  and its variance is given by

$$V(e_H(Y)) = \sum_{i \in U_A} \sigma_{A_i}^2 \tilde{w}_{H_i}^2 \pi_i^A + \sum_{i \in U_B} \sigma_{B_i}^2 \tilde{w}_{H_i}^2 \pi_i^B +$$

$$\sum_{i,j \in U_A} y_i y_j (\tilde{w}_{H_i} \tilde{w}_{H_j} \pi_{ij}^A - 1) + \sum_{i,j \in U_B} y_i y_j (\tilde{w}_{H_i} \tilde{w}_{H_j} \pi_{ij}^B - 1) \tag{11}$$

where

$$\tilde{w}_{H_i} = \begin{cases} w_{A_i} & \text{if } i \in a \\ w_{B_i} & \text{if } i \in b \\ \eta w_{A_i} + (1 - \eta) w_{B_i} & \text{if } i \in ab \end{cases} \tag{12}$$

Proof.

Observe that estimator  $e_H(Y)$  can be rewritten as

$$e_H(Y) = \sum_{i \in s_A} \tilde{w}_{H_i} R_{A_i} + \sum_{i \in s_B} \tilde{w}_{H_i} R_{B_i}.$$

Thus  $E(e_H(Y))$  can be calculated as in Theorem 1 by changing  $\tilde{w}_{H_i}$  by  $\tilde{w}_{SF_i}$ .

Respect to the variance, the proof is the same that in Theorem 2 by replacing  $\tilde{w}_{H_i}$  with  $\tilde{w}_{SF_i}$ . The variance estimators for  $V(e_H(Y))$  are obtained following the procedure described in the previous section.

### 6.1 Selecting the weight for the average

The choice of weight  $\eta$  is an important issue in dual-frame estimators because the efficiency of the estimator depends on this value.

[9] proposed choosing  $\eta$  to minimize the variance of the estimator. Thus, by minimizing (11) with respect to  $\eta$  and after some algebraic calculus we obtain the value

$$\eta_o = \frac{2 * A_6 - A_2 + A_3 + 2A_5}{2A_1 + 2A_4 + 2A_5}$$

where

$$\begin{aligned}
 A_1 &= \sum_{i \in U_{ab}} \sigma_{A_i}^2 \tilde{w}_{H_i}^2 \pi_i^A - \sum_{i \in U_{ab}} \sigma_{B_i}^2 \tilde{w}_{H_i}^2 \pi_i^B, \\
 A_2 &= \sum_{i \in U_a, j \in U_{ab}} y_i y_j (\tilde{w}_{H_i} \tilde{w}_{H_j} \pi_{ij}^A), \\
 A_3 &= \sum_{i \in U_b, j \in U_{ab}} y_i y_j (\tilde{w}_{H_i} \tilde{w}_{H_j} \pi_{ij}^B), \\
 A_4 &= \sum_{i, j \in U_{ab}} y_i y_j (\tilde{w}_{H_i} \tilde{w}_{H_j} \pi_{ij}^A), \\
 A_5 &= \sum_{i, j \in U_{ab}} y_i y_j (\tilde{w}_{H_i} \tilde{w}_{H_j} \pi_{ij}^B) \text{ and} \\
 A_6 &= \sum_{i \in U_{ab}} \sigma_{B_i}^2 \tilde{w}_{H_i}^2 \pi_i^B.
 \end{aligned}$$

Thus, the optimal estimator obtained with  $\eta_o$  is a function of the variances and covariances of the estimated domain totals and then the optimal estimates will differ for different response variables, leading to internal inconsistency. In practice, the values  $A_j, j = 1, \dots, 6$  are unknown, and so the optimal value of  $\eta$  cannot be calculated and must be estimated from the sample data. This is computationally complex and also affects optimality since the extra variability in estimating the variance leads to larger mean square errors. [34] used  $\eta = 1/2$  in their study of a dual-frame survey in which frame A was a landline telephone frame and frame B was a cell-phone frame. For this purpose, the value of  $\eta = 1/2$  is frequently recommended (see, for example, [35]).

It is also possible to estimate  $\eta$  using

$$\hat{\eta}_N = N_a N_B v(\hat{N}_{ba}) / [N_b N_A v(\hat{N}_{ab}) + N_a N_B v(\hat{N}_{ba})], \quad (13)$$

(see [36]) where  $\hat{N}_{ab} = \sum_{i \in s_{ab}} w_{A_i}$  is the Horvitz-Thompson estimator for the size of domain  $ab$  from the design in frame A and  $\hat{N}_{ba} = \sum_{i \in s_{ba}} w_{B_i}$  is the Horvitz-Thompson estimator for the size of domain  $ba$  from the design in frame B. In this case, it does not depend on the values of the main variable.

### 7 Simulation study

We conducted a simulation study to analyse the performance of the proposed estimators for surveys from two-frame finite populations. Our simulations are programmed in R.

The simulated population has the dimension  $N = 2350$ . The values of the variable of interest  $y$  are generated by two forms. One of them is from a normal distribution  $y_i \sim N(5000, 500)$ , for  $i = 1, \dots, 2350$ , i.e., for quantitative models, and the other is from binomial distribution  $y_i \sim Bi(2350, 0.5)$ , for  $i = 1, \dots, 2350$ , qualitative models. Units are randomly assigned to the two frames, A and B, according to three different scenarios depending on the overlap domain size  $N_{ab}$ . The

first scenario has a *small* overlap domain size and units are assigned to domain  $a, b$  or  $ab$  depending on the values taken by a binomial random variable  $g_i \sim Bi(2, 0.3)$ . In particular, if  $g_i = 0$  then  $i \in a$ , if  $g_i = 1$  then  $i \in b$  and if  $g_i = 2$  then  $i \in ab$ . In the quantitative models, the resulting sizes of the two frames are  $N_A=1181$  and  $N_B=934$  and, consequently, the overlap domain size is  $N_{ab}=235$ . In the qualitative models, they are  $N_A=1133$ ,  $N_B=1006$  and  $N_{ab}=211$ .

The second and the third scenarios have *large* and *medium* overlap domain sizes, respectively, depending on the values of  $g_i \sim Bi(2, 0.5)$ , but units are assigned to each domain in different ways in each scenario. In particular, we have 0 for domain  $a$ , 1 for domain  $ab$  and 2 for domain  $b$  in the second scenario and 0 for domain  $b$ , 1 for domain  $a$  and 2 for domain  $ab$  in the third scenario. For the quantitative models, the resulting frame sizes in the second scenario are given by  $N_A=561$  and  $N_B=606$  and the overlap domain size is  $N_{ab}=1183$ , while for the third scenario we have  $N_A=1183$ ,  $N_B=561$  and  $N_{ab}=606$ . In the qualitative models in the second scenario, they are  $N_A=578$ ,  $N_B=602$  and  $N_{ab}=1170$ , while for the third scenario they are  $N_A=1170$ ,  $N_B=578$  and  $N_{ab}=602$ .

In the quantitative models, the units from frames A and B are then divided for each scenario into six strata as follows:

- *Small*,  $N_h^A = (599, 284, 92, 158, 86, 197)$ ,  $N_h^B = (510, 261, 73, 120, 75, 130)$
- *Large*,  $N_h^A = (729, 364, 116, 183, 116, 236)$ ,  $N_h^B = (775, 384, 113, 190, 111, 216)$
- *Medium*,  $N_h^A = (775, 384, 113, 190, 111, 216)$ ,  $N_h^B = (496, 252, 71, 127, 73, 148)$ .

And in the qualitative models they are:

- *Small*,  $N_h^A = (569, 292, 81, 147, 84, 171)$ ,  $N_h^B = (520, 253, 81, 124, 82, 157)$
- *Large*,  $N_h^A = (751, 376, 108, 176, 108, 229)$ ,  $N_h^B = (752, 368, 110, 191, 118, 233)$
- *Medium*,  $N_h^A = (752, 368, 110, 191, 118, 233)$ ,  $N_h^B = (497, 256, 82, 133, 74, 138)$ .

Samples from frames A and B are selected using stratified simple random sampling. For each scenario, we draw four different combinations of sample sizes for frame A and frame B, which correspond to the following numbers of units per stratum:

- c1:  $n_A = (15, 20, 15, 20, 15, 20) = 105$ ,  $n_B = (25, 20, 25, 20, 25, 20) = 135$ ,
- c2:  $n_A = (30, 40, 30, 40, 30, 40) = 210$ ,  $n_B = (25, 20, 25, 20, 25, 20) = 135$ ,
- c3:  $n_A = (15, 20, 15, 20, 15, 20) = 105$ ,  $n_B = (50, 40, 50, 40, 50, 40) = 270$ ,
- c4:  $n_A = (30, 40, 30, 40, 30, 40) = 210$ ,  $n_B = (50, 40, 50, 40, 50, 40) = 270$ .

For each sample in each scenario, we computed point estimators for three quantitative and qualitative models in frame B. In frame A, we assume direct questionnaire answering. In the quantitative models, we compare the

**Table 1:** Relative mean squared error (*relative bias*) of the estimators compared. Single-frame approach (SF) and Dual-frame approach (DF) for selection of the weighting parameter  $\eta$ . Randomized response models for the quantitative variables: Eichhorn and Hayre, BBB and FQRR

	<i>Small</i>				<i>Large</i>				<i>Medium</i>			
	c1	c2	c3	c4	c1	c2	c3	c4	c1	c2	c3	c4
EICHHORN AND HAYRE												
SF	0.10 <i>0.00</i>	0.08 <i>-0.08</i>	0.09 <i>0.08</i>	0.05 <i>0.11</i>	0.18 <i>-0.08</i>	0.13 <i>-0.12</i>	0.16 <i>0.10</i>	0.09 <i>0.03</i>	0.16 <i>-0.00</i>	0.08 <i>-0.16</i>	0.17 <i>0.12</i>	0.08 <i>0.08</i>
DF $\eta_{1/2}$	0.09 <i>-0.03</i>	0.08 <i>-0.09</i>	0.06 <i>0.03</i>	0.04 <i>0.12</i>	0.17 <i>-0.08</i>	0.13 <i>-0.12</i>	0.11 <i>0.05</i>	0.08 <i>0.06</i>	0.12 <i>-0.01</i>	0.08 <i>-0.17</i>	0.08 <i>0.04</i>	0.05 <i>0.06</i>
DF $\eta_{opt}$	0.10 <i>-0.43</i>	0.08 <i>-0.37</i>	0.08 <i>-0.22</i>	0.05 <i>-0.11</i>	0.16 <i>-0.07</i>	0.14 <i>-0.18</i>	0.13 <i>0.15</i>	0.08 <i>0.06</i>	0.16 <i>-0.23</i>	0.08 <i>-0.37</i>	0.17 <i>-0.01</i>	0.08 <i>-0.02</i>
DF $\eta_N$	0.10 <i>-0.32</i>	0.08 <i>-0.30</i>	0.08 <i>-0.14</i>	0.05 <i>-0.05</i>	0.17 <i>0.21</i>	0.13 <i>0.03</i>	0.16 <i>0.27</i>	0.09 <i>0.17</i>	0.18 <i>-0.08</i>	0.09 <i>-0.23</i>	0.19 <i>0.06</i>	0.09 <i>0.05</i>
BBB												
SF	0.06 <i>0.03</i>	0.05 <i>0.01</i>	0.07 <i>0.05</i>	0.03 <i>0.12</i>	0.13 <i>-0.05</i>	0.10 <i>-0.02</i>	0.13 <i>0.05</i>	0.06 <i>0.05</i>	0.14 <i>0.00</i>	0.06 <i>-0.08</i>	0.15 <i>0.08</i>	0.06 <i>0.11</i>
DF $\eta_{1/2}$	0.06 <i>-0.00</i>	0.05 <i>0.00</i>	0.04 <i>0.01</i>	0.03 <i>0.13</i>	0.12 <i>-0.04</i>	0.09 <i>-0.00</i>	0.08 <i>0.01</i>	0.06 <i>0.07</i>	0.09 <i>-0.00</i>	0.06 <i>-0.09</i>	0.07 <i>0.01</i>	0.04 <i>0.08</i>
DF $\eta_{opt}$	0.06 <i>-0.35</i>	0.05 <i>-0.26</i>	0.06 <i>-0.22</i>	0.03 <i>-0.07</i>	0.11 <i>0.07</i>	0.10 <i>-0.03</i>	0.11 <i>0.14</i>	0.06 <i>0.11</i>	0.14 <i>-0.15</i>	0.06 <i>-0.22</i>	0.16 <i>-0.02</i>	0.07 <i>0.03</i>
DF $\eta_N$	0.06 <i>-0.28</i>	0.05 <i>-0.21</i>	0.07 <i>-0.18</i>	0.03 <i>-0.03</i>	0.12 <i>0.23</i>	0.10 <i>0.13</i>	0.13 <i>0.22</i>	0.06 <i>0.19</i>	0.15 <i>-0.07</i>	0.07 <i>-0.14</i>	0.17 <i>0.01</i>	0.07 <i>0.08</i>
FQRR												
SF	0.10 <i>0.00</i>	0.08 <i>0.06</i>	0.09 <i>0.10</i>	0.05 <i>0.17</i>	0.18 <i>0.08</i>	0.13 <i>0.03</i>	0.17 <i>0.13</i>	0.09 <i>0.11</i>	0.16 <i>0.12</i>	0.09 <i>-0.06</i>	0.17 <i>0.14</i>	0.08 <i>0.16</i>
DF $\eta_{1/2}$	0.09 <i>-0.03</i>	0.08 <i>0.06</i>	0.06 <i>0.06</i>	0.04 <i>0.17</i>	0.16 <i>0.08</i>	0.14 <i>0.04</i>	0.11 <i>0.07</i>	0.08 <i>0.13</i>	0.11 <i>0.10</i>	0.08 <i>-0.07</i>	0.08 <i>0.07</i>	0.05 <i>0.13</i>
DF $\eta_{opt}$	0.09 <i>-0.43</i>	0.09 <i>-0.22</i>	0.08 <i>-0.20</i>	0.05 <i>-0.04</i>	0.15 <i>0.14</i>	0.14 <i>-0.02</i>	0.13 <i>0.19</i>	0.08 <i>0.16</i>	0.16 <i>-0.08</i>	0.08 <i>-0.24</i>	0.17 <i>0.00</i>	0.07 <i>0.05</i>
DF $\eta_N$	0.09 <i>-0.32</i>	0.08 <i>-0.15</i>	0.09 <i>-0.12</i>	0.05 <i>0.01</i>	0.17 <i>0.36</i>	0.13 <i>0.18</i>	0.16 <i>0.30</i>	0.09 <i>0.26</i>	0.18 <i>0.04</i>	0.09 <i>-0.12</i>	0.19 <i>0.07</i>	0.09 <i>0.13</i>

Eichhorn and Hayre model, the BBB model with  $p_1 = 0.6$  and  $p_2 = 1 - p_1$ , and the FQRR model with  $p_1 = 0.3$ ,  $p_2 = 0.4$  and  $p_3 = 0.1$ .

For the distribution of the scramble variable, we follow [21] and select a  $F_{20,20}$  distribution. In the qualitative models, we compare the Warner model with  $p = 0.6$ , the H model with  $p_1 = 0.6$  and  $p_2 = p_3 = (1 - p_1)/2$ , and the U model with  $p_1 = 0.6$  and  $p_2 = 1 - p_1$ .

For each estimator  $\hat{Y}$  of the population total  $Y$ , we computed the relative bias  $RB = E_{MC}(\hat{Y} - Y)/Y * 100\%$  (as a percentage) and the relative mean squared error  $RMSE = E_{MC}[(\hat{Y} - Y)^2]/Y * 100\%$  (as a percentage), where  $E_{MC}$  denotes the average based on 1000 simulation runs.

Tables 1 and 2 show our results, from which some important conclusions can be drawn:

1. For all models with quantitative variables, the relative bias is less than 0.5% for all sample sizes and for all scenarios.
2. For the H and U models, the estimator with  $\eta_{opt}$  is more biased than the other estimators, but in any case, it is less than 1%. For the Warner model, the relative

bias exceptionally reaches 2% (See. *Small*,  $n_A = 105$ ,  $n_B = 135$ ) and 1% (See. *Large*,  $n_A = 105$ ,  $n_B = 135$ ).

3. For the same  $p_1$  value in the random mechanism, the least efficient (highest relative mean square error) model is the Warner model. The estimated mean squared errors decrease drastically when the H and U models are used.
4. In general, the smallest relative mean squared errors are obtained in the *small* scenario, both for the quantitative and the qualitative variables.
5. For all models with quantitative variables, the best results in terms of efficiency are achieved with the dual frame estimator with  $\eta_{1/2}$  (except for only three cases over 36 situations).
6. For the H and U models, the best results in terms of efficiency are achieved with the dual frame estimator with  $\eta_{1/2}$  (except for only two cases over 24 situations). For the Warner model, the most efficient estimator is the dual frame with  $\eta_{opt}$ .



**Table 2:** Relative mean squared error (*relative bias*) of the estimators compared. Single-frame approach (SF) and Dual-frame approach (DF) for selection of the weighting parameter  $\eta$ . Randomized response models for the qualitative variables: Warner, H and U models

	<i>Small</i>				<i>Large</i>				<i>Medium</i>			
	c1	c2	c3	c4	c1	c2	c3	c4	c1	c2	c3	c4
WARNER												
SF	6.90	6.10	3.91	3.54	9.49	6.75	5.75	4.55	5.56	4.34	3.24	2.79
	-0.35	2.05	-0.75	0.09	1.26	-0.35	0.28	-0.40	-0.50	-0.32	0.09	0.71
DF $\eta_{1/2}$	6.76	6.16	3.69	3.47	8.50	7.44	4.27	4.07	4.92	4.30	2.56	2.44
	-0.32	2.05	-0.80	0.10	1.13	-0.25	0.16	-0.27	-0.35	-0.26	0.04	0.53
DF $\eta_{opt}$	6.55	5.93	3.55	3.32	6.36	5.22	3.34	2.90	4.32	3.55	2.32	2.13
	-0.51	2.07	-1.25	-0.07	0.79	-0.24	-0.10	0.30	-0.45	-0.10	-0.51	0.02
DF $\eta_N$	6.90	6.10	3.91	3.54	9.61	6.78	5.87	4.62	5.92	4.65	3.41	2.99
	-0.55	1.93	-0.97	-0.05	1.56	-0.09	0.36	-0.32	-0.66	-0.37	0.05	0.71
H MODEL												
SF	1.25	0.91	0.84	0.58	1.39	0.97	1.01	0.67	1.24	0.78	0.94	0.63
	0.10	-0.15	-0.06	-0.14	-0.25	-0.45	0.02	0.45	-0.01	-0.14	0.02	0.22
DF $\eta_{1/2}$	1.23	0.91	0.81	0.58	1.32	1.04	0.93	0.64	1.17	0.77	0.87	0.59
	0.13	-0.15	-0.07	-0.14	-0.22	-0.41	-0.04	0.46	-0.04	-0.15	0.08	0.13
DF $\eta_{opt}$	1.23	0.91	0.82	0.58	1.29	0.93	0.95	0.62	1.19	0.76	0.92	0.59
	-0.64	-0.63	-0.79	-0.59	-0.70	-0.70	-0.39	0.24	-0.89	-0.60	-0.60	-0.27
DF $\eta_N$	1.25	0.91	0.84	0.58	1.41	0.98	1.02	0.68	1.26	0.80	0.96	0.65
	-0.17	-0.31	-0.26	-0.28	0.04	-0.26	0.16	0.58	-0.12	-0.23	-0.03	0.19
U MODEL												
SF	0.96	0.75	0.71	0.48	1.14	0.81	0.80	0.57	1.09	0.73	0.85	0.54
	-0.07	0.33	-0.13	-0.25	0.00	-0.23	0.40	0.22	0.10	0.21	0.17	0.14
DF $\eta_{1/2}$	0.95	0.75	0.70	0.48	1.08	0.83	0.74	0.55	1.04	0.72	0.81	0.52
	-0.02	0.33	-0.13	-0.25	-0.01	-0.18	0.25	0.25	0.14	0.21	0.17	0.08
DF $\eta_{opt}$	0.98	0.76	0.71	0.48	1.10	0.82	0.76	0.55	1.08	0.73	0.84	0.53
	-0.97	-0.32	-0.87	-0.73	-0.83	-0.73	-0.16	-0.06	-0.93	-0.50	-0.49	-0.36
DF $\eta_N$	0.96	0.75	0.71	0.48	1.15	0.80	0.81	0.58	1.11	0.75	0.87	0.55
	-0.32	0.18	-0.33	-0.38	0.27	-0.04	0.53	0.32	-0.03	0.12	0.12	0.11

## 8 Conclusions

Social science researchers are increasingly examining sensitive issues such as drug use, sexual orientation and lifestyle, race relations, abortion and illegal activities. At the same time, the public demands privacy and protection and has become highly suspicious of intruders into their lives.

One way to reduce response bias in self-report methodology is to use the randomized response technique, which may provide more valid data than traditional methods, by giving the respondents more privacy when the information requested is very sensitive.

In this paper, we present a new procedure aimed at determining a population total using a randomized-response model when data are obtained from two frames. We introduce different ways of combining estimates from the different frames. In practice, a different sampling procedure might feasibly be applied for each frame, or even no randomization at all (i.e., direct response) for a particular frame. The use of RR techniques has advantages but also drawbacks (the variance of estimates is increased by the randomization and individual response patterns cannot be interpreted

directly, due to the observation of randomized responses, nor can individuals or groups of individuals be compared). Nevertheless, by making combined use of RR and direct answering in the sample, information that is both more valid and more reliable can be obtained.

A broad range of randomized response models can be applied in a survey with two frames, and the proposed approach enables us to address these situations.

This paper considers two estimators that combine the information obtained by the randomized schemes in each frame. These estimators are based on the estimators proposed by [33] and by [31] but a wide variety of estimators have been reported in the literature on multiple frames, according to two main approaches: single-frame and dual-frame ([37] and [38]). See [35] for a good review of their properties. According to this author, all these estimators can be expressed as a linear combination of  $y$  values for convenient weights  $\tilde{w}_i, i \in U$ . Consequently, these estimators can be used to define new RR estimators.

In this study, for the sake of clarity, only two frames were used. The proposed method could also be extended to three or more frames by using the method suggested by [39].

## Acknowledgements

This study was partially supported by Ministerio de Educación y Ciencia (grant MTM2012-35650, Spain) and by Consejería de Economía, Innovación, Ciencia y Empleo (grant SEJ2954, Junta de Andalucía).

## References

- [1] I. Krumpal, Determinants of social desirability bias in sensitive surveys: a literature review, *Quality & Quantity* **47**, (2013) 2025–2047.
- [2] S. L. Warner, Randomized response: A survey technique for eliminating evasive answer bias, *Journal of the American Statistical Association* **60**, 63–69.
- [3] E. Coutts, B. Jann, Sensitive questions in online surveys: experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT), *Sociol. Methods Res.* **40**, (2011) 169–193.
- [4] G. J. L. M. Lensvelt-Mulders, J. J. Hox, P. G. M. van der Heijden, C. J. M. Maas, Meta-analysis of randomized response research: thirty-five years of validation, *Sociol. Methods Res.* **33**, (2005) 319–348.
- [5] R. Arnab, Optional randomized response techniques for complex survey designs, *Biom. J.* **46**, (2004) 114–124.
- [6] C. N. Bouza, C. Herrera, P. G. Mitra, A review of randomized responses procedures: the qualitative variable case, *Investigación Oper.* **31**, (2010) 240–247.
- [7] A. Chaudhuri, *Randomized response and indirect questioning techniques in surveys*, Statistics: Textbooks and Monographs, CRC Press, Boca Raton, FL, 2011.
- [8] M. Winglee, I. Park, K. Rust, B. Liu, G. Shapiro, A case study in dual-frame estimation methods, *Proceedings of the Survey Research Methods Section, ASA* (2007) 3233–3238.
- [9] H. O. Hartley, Multiple frame methodology and selected applications, *Sankhyā C.* **36**, (1974) 99–118.
- [10] A. van den Hout, U. Böckenholt, P. G. M. van der Heijden, Estimating the prevalence of sensitive behaviour and cheating with a dual design for direct questioning and randomized response, *J. R. Stat. Soc. Ser. C. Appl. Stat.* **59**, (2010) 723–736.
- [11] D. G. Horvitz, W. R. Simmons, The unrelated question randomized response model, *Proceedings of the Survey Research Methods Section, ASA* (1967) 65–72.
- [12] B. G. Greenberg, A.-L. A. Abul-Ela, W. R. Simmons, D. G. Horvitz, The unrelated question randomized response model: Theoretical framework, *Journal of the American Statistical Association* **64**, 520–539.
- [13] D. G. Horvitz, B. G. Greenberg, J. R. Abernathy, Randomized response: A data-gathering device for sensitive questions, *International Statistical Review* **44**, (1976) 181–196.
- [14] N. S. Mangat, R. Singh, An alternative randomized response procedure **77**, (1990) 439–442.
- [15] J. L. Devore, A note on the randomized response technique, *Communications in Statistics - Theory and Methods* **6**, (1977) 1525–1529.
- [16] D. Tracy, N. Mangat, Some development in randomized response sampling during the last decade—a follow up of review by Chaudhuri and Mukerjee, *Journal of Applied Statistical Science* **4**, (1996) 533–544.
- [17] R. Arnab, Optimum sampling strategies under randomized response surveys, *Biom. J.* **44**, (2002) 490–495.
- [18] H.-J. Chang, C.-L. Wang, K.-C. Huang, Using randomized response to estimate the proportion and truthful reporting probability in a dichotomous finite population, *J. Appl. Stat.* **31**, (2004) 565–573.
- [19] J.-P. Fox, C. Wyrick, A mixed effects randomized item response model, *Journal of Educational and Behavioral Statistics* **33**, (2008) 389–415.
- [20] B. G. Greenberg, R. R. Kuebler, J. R. Abernathy, D. G. Horvitz, Respondent hazards in the unrelated question randomized response model, *J. Stat. Plann. Inference* **1**, (1977) 53–60.
- [21] B. H. Eichhorn, L. S. Hayre, Scrambled randomized response methods for obtaining sensitive quantitative data, *Journal of Statistical Planning and Inference* **7**, (1983) 307–316.
- [22] S. K. Bar-Lev, E. Bobovitch, B. Boukai, A note on randomized response models for quantitative data, *Metrika* **60**, (2004) 255–260.
- [23] C. R. Gjestvang, S. Singh, An improved randomized response model: estimation of mean, *J. Appl. Stat.* **36**, (2009) 1361–1367.
- [24] S. A. Eriksson, A new model for randomized response, *International Statistical Review / Revue Internationale de Statistique* **41**, 101–113.
- [25] A. Saha, A simple randomized response technique in complex surveys, *Metron - International Journal of Statistics* **LXV**, (2007) 59–66.
- [26] G. Diana, P. F. Perri, A calibration-based approach to sensitive data: a simulation study, *Journal of Applied Statistics* **39**, (2012) 53–65.
- [27] A. Chaudhuri, R. Mukherjee, *Randomized response. Theory and techniques.*, Statistics: Textbooks and Monographs, 85. New York etc.: Marcel Dekker, Inc. xvi, 162 p. , 1988.
- [28] R. Arnab, Randomized response trials: a unified approach for qualitative data, *Comm. Statist. Theory Methods* **25**, (1996) 1173–1183.
- [29] V. Soberanis-Cruz, G. Ramírez-Valverde, S. Pérez-Elizalde, F. González-Cossio, Muestreo de respuestas aleatorizadas en poblaciones finitas: un enfoque unificador, *Agrociencia*.
- [30] M. Rueda, A. Arcos, S. Singh, A generalized approach to randomised response for quantitative variables, *Quality & Quantity* (2014) (in press)
- [31] M. D. Bankier, Estimators based on several stratified samples with applications to multiple frame surveys, *Journal of the American Statistical Association* **81**, (1986) 1074–1079.
- [32] K. Wolter, *Introduction to Variance estimation*, Springer-Verlag, New York, 2003.
- [33] H. O. Hartley, Multiple frame surveys, in: *Proceedings of the Social Statistics Section, American Statistical Association*, 1962, 203–206.
- [34] J. M. Brick, W. S. Edwards, S. Lee, Sampling telephone numbers and adults, interview length, and weighting in the california health interview survey cell phone pilot study **71**, (2007) 793–813.
- [35] F. Mecatti, A single frame multiplicity estimator for multiple frame surveys, *Survey methodology* **33**, (2007) 151–157.
- [36] S. L. Lohr, J. N. K. Rao, Inference from dual frame surveys, *Journal of the American Statistical Association* **95**, (2000) 271–280.

- [37] W. A. Fuller, L. F. Burmeister, Estimators for samples selected from two overlapping frames, Proceedings of social science section of The American Statistical Association.
- [38] C. J. Skinner, J. N. K. Rao, Estimation in dual frame surveys with complex designs, Journal of the American Statistical Association **91**, (1996) 349–356.
- [39] A. C. Singh, F. Mecatti, Generalized multiplicity-adjusted horvitz-thompson estimation as a unified approach to multiple frame surveys, Journal of official statistics **27**.



**Beatriz Cobo** is PhD student with a scholarship from Science and Innovation Ministry, Spain.



**Maria del Mar Rueda**, PhD is Full Professor at the Department of Statistics and Operational Research, University of Granada. Her professional interests include teaching and study of survey research methods, with particular emphasis on the use of auxiliary information

in complex survey designs, methods for the analysis of survey data and survey nonresponse.



**Antonio Arcos**, PhD is a Professor at the Department of Statistics and Operational Research, University of Granada. He specializes in the design and analysis of complex surveys. His research has focused on using auxiliary information in surveys in order to improve

the accuracy of survey estimates.