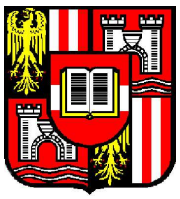

Department for Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2007-23

Comparing the efficiency of randomized response techniques under uniform conditions

Andreas Quatember

May 2007

Abstract

In this report it is shown, that one-stage questioning designs cannot be less efficient than their two-stage versions when we take control of the level of the respondents' privacy protection provided by the designs. Furthermore the calculation of limits of minimum privacy protection – done herein by the Leysieffer-Warner measures – allows to distinguish between strategies, that are useable for all sensitive subjects and others, which should only be applied, if the possession but not the nonpossession of some attribute is embarrassing.

1 Introduction

The pioneering work in the field of randomized response strategies was written by Warner (1965). Let U be the universe of N population units and U_A be the set of N_A elements from U, that belong to a class A of a sensitive categorial variable under study. Furthermore let U_{A^c} be the group of N_{A^c} elements, that do not belong to this class ($N = N_A + N_{A^c}$, $U = U_A \cup U_{A^c}$). Let the parameter of interest be

$$\pi_A = \frac{N_A}{N}, \tag{1}$$

which is the relative size of U_A. In Warner's questioning design (W) each respondent has to answer randomly either with probability p (for instance realized by drawing a card) the question "Are you a member of group U_A?" or with probability $1-p$ the alternative question "Are you a member of group U_{A^C}?". p is the *design parameter* of the Warner technique.

Assuming that the randomized questioning will ensure the cooperation of all selected sample units as well as truthful responses Warner considered the following unbiased estimator of π_A:

$$\hat{\pi}_A^W = \frac{\hat{\pi}_y + p - 1}{2p - 1} \tag{2}$$

(for $p \neq 0,5$) with $\hat{\pi}_y$, the proportion of "yes"-answers in the sample of size n (Warner (1965), p.65) and estimator of π_y, the probability of such an answer.

The variance of the Warner estimator (2) for simple random sampling without replacement (wor) is

$$V_{wor}(\hat{\pi}_A^W) = \frac{\pi_A \cdot (1 - \pi_A)}{n} \cdot \frac{N - n}{N - 1} + \frac{p \cdot (1 - p)}{n \cdot (2p - 1)^2} \tag{3}$$

(Kim and Flueck (1978), p.347). For $n \to N$ this variance – in contrary to the variance of the estimator $\hat{\pi}_A^{dir} = \hat{\pi}_y$ for the direct questioning method (dir) with the unappropriate assumption of fullresponse – does not approach zero but

$$\lim_{n \to N} V_{wor}(\hat{\pi}_A^W) = \frac{p \cdot (1 - p)}{N \cdot (2p - 1)^2}.$$

For simple random sampling with replacement (wr) or for large populations (3) reduces to

$$V_{wr}(\hat{\pi}_A^W) = \frac{\pi_A \cdot (1 - \pi_A)}{n} + \frac{p \cdot (1 - p)}{n \cdot (2p - 1)^2} \tag{4}$$

(Warner (1965), p.65). It is the right one of the two summands in both (3) and (4) that corresponds to the increase of the variance caused by the use of Warner's technique instead of direct questioning.

Obviously these variances get larger the closer p is chosen to 0.5. But at the same time a design parameter next to 0.5 ensures a high level of the individual's confidence in the protection of his or her privacy. These opposite interests have both to be taken into account when the value for the design parameter p is to be fixed. Therefore it would be desirable to determine an optimum value $p = p_{opt}$ that far away from 0.5 that the respondents' willingness to cooperate will just be guaranteed. Greenberg et al. (1969) suggested p_{opt} around 0.8 or 0.2 as practicable for this purpose (p.526). But certainly the value of p_{opt} will depend on the subject of interest, meaning that the more sensitive a subject is, the closer to 0.5 the value for p_{opt} has to be chosen. So the practical experience of the user in this field together with empirical studies will surely help in determining p_{opt}.

If $p = 1$ (or $p = 0$), the interviewee is directly asked about his or her membership in U_A or U_{A^c}. The direct questioning method therefore can be seen as a special case of Warner's questioning strategy. Caused by the total loss of his or her privacy it results in the greatest respondent's burden leading to the highest nonresponse or untruthful answer rate for the item under investigation.

Example 1: Let N be large, $n = 100$ and $\pi_A = 0.2$. Furthermore let the optimum design parameter of Warner's questioning design for a certain sensitive variable be $p_{opt} = 0.8$. (These parameters will be used all throughout the examples to follow.) Then for simple random sampling with or without replacement the estimator (2) has a standard deviation of $7.775 \cdot 10^{-2}$. With the unappropriate assumption of fullresponse for the estimator $\hat{\pi}_A^{dir}$ obtained by the direct method we would have: $[V(\hat{\pi}_A^{dir})]^{1/2} = 4 \cdot 10^{-2}$. $\triangle$

A generalised two-stage version $W2$ of strategy W, in which the randomization of questions takes place in two steps, can be described as follows: At stage I the sample unit is asked with probability p_I: "Are you a member of group U_A?", whereas he or she is referred to stage II with probability $1 - p_I$. At this stage Warner's one-stage technique with design parameter p_{II} is used. As the two-stage strategy is nothing else than the one-stage questioning design with design parameter $p = p_I + (1-p_I) \cdot p_{II}$ the unbiased estimator for π_A is:

$$\hat{\pi}_A^{W2} = \frac{\hat{\pi}_y - (1 - p_I) \cdot (1 - p_{II})}{p_{II} + (2p_I - 1) \cdot (1 - p_{II})} \tag{5}$$

(for $p_I + (1 - p_I) \cdot p_{II} \neq 0.5$) and the variances of $\hat{\pi}_A^{W2}$ for simple random sampling without and with replacement can be calculated by inserting $p_I + (1-p_I) \cdot p_{II}$ instead of p in (3) and (4) .

Mangat and Singh (1990) considered this method for the special case of $p_{II} = p$ with p being the design parameter of the Warner one-stage questioning design. They have shown, that the unbiased estimator

$$\hat{\pi}_A^{W2*} = \frac{\hat{\pi}_y - (1 - p_I) \cdot (1 - p)}{p + (2p_I - 1) \cdot (1 - p)} \tag{6}$$

$(p_I + (1 - p_I) \cdot p \neq 0.5)$ for π_A would have a smaller variance than $\widehat{\pi}_A^W$, if $p_I > (1 - 2p)/(1 - p)$ (p.440). In the generalised two-stage Warner design this can be formulated in the following way: The estimator $\widehat{\pi}_A^{W2}$ would be more accurate than $\widehat{\pi}_A^W$, when we would choose $p_I + (1 - p_I) \cdot p_{II} > max(p, 1 - p)$. This means that the two-stage Warner strategy would always be more efficient than the one-stage technique, when we choose $p_I + (1 - p_I) \cdot p_{II}$ farther away from 0.5 than p.

Example 2: For a two-stage Mangat-Singh estimator (6) let the design parameters be $p_I = 0.2$ and $p_{II} = p = 0.8$, so that $p_I + (1 - p_I) \cdot p_{II} = 0.84$. The theoretical standard deviation of the estimator for simple random sampling without replacement would then be $6.713 \cdot 10^{-2}$.$\triangle$

But this is not a great success, because if Warner's design parameter p was chosen optimally ($p = p_{opt}$) in the sense described above, then $p_I + (1 - p_I) \cdot p_{II}$ simply *should not* be chosen in this way, because a choice of $p_I + (1 - p_I) \cdot p_{II} > p_{opt}$ would then automatically cause nonresponse and/or untruthful answers. This would set us back to the starting point of our problem.

2 To measure the respondent's privacy protection

To apply these considerations also to other randomized response techniques we have to look for measures of protection of the respondent's privacy: Let

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

Then the following ratios ρ_1 and ρ_0 of a posteriori probabilities may measure privacy protection with respect to the respondent's answer:

$$\rho_1 = \frac{P(i \in U_A | y_i = 1)}{P(i \in U_{A^c} | y_i = 1)} \quad \text{and} \quad \rho_0 = \frac{P(i \in U_{A^c} | y_i = 0)}{P(i \in U_A | y_i = 0)}.$$

A totally protected privacy would result in

$$\rho_{1,tot} = \frac{P(i \in U_A)}{P(i \in U_{A^c})} = \frac{\pi_A}{1 - \pi_A} \quad \text{and} \quad \rho_{0,tot} = \frac{P(i \in U_{A^c})}{P(i \in U_A)} = \frac{1 - \pi_A}{\pi_A}.$$

In such cases the concrete answer provides absolutely no usable information with respect to the respondent's possession or nonpossession of the attribute A. The more ρ_1 or ρ_0 differ from $\rho_{1,tot}$ or $\rho_{0,tot}$, the lower the privacy of the interviewee is protected, given his or her answer. For the direct questioning design, offering absolutely no such protection, these measures are $\rho_1^{dir} = \rho_0^{dir} = \infty$ for $p = 1$ and $\rho_1^{dir} = \rho_0^{dir} = 0$ for $p = 0$ respectively.

Because of

$$P(i \in U_A | y_i) \cdot P(y_i) = P(y_i | i \in U_A) \cdot P(i \in U_A),$$

which also applies to $i \in U_{A^c}$, the measure ρ_1 can be rewritten as

$$\rho_1 = \frac{\pi_A}{1 - \pi_A} \cdot \lambda_1$$

with

$$\lambda_1 = \frac{P(y_i = 1 | i \in U_A)}{P(y_i = 1 | i \in U_{A^c})}. \tag{7}$$

Consequently ρ_0 is given by

$$\rho_0 = \frac{1 - \pi_A}{\pi_A} \cdot \lambda_0$$

with

$$\lambda_0 = \frac{P(y_i = 0 | i \in U_{A^c})}{P(y_i = 0 | i \in U_A)}. \tag{8}$$

The ratios λ_1 and λ_0 of a priori probabilities have been suggested by Leysieffer and Warner (1976) as "measures of jeopardy" (p.650). For a total protected privacy (7) and (8) would result in

$$\lambda_{1,tot} = \rho_{1,tot} \cdot \frac{1 - \pi_A}{\pi_A} = 1$$

and

$$\lambda_{0,tot} = \rho_{0,tot} \cdot \frac{\pi_A}{1 - \pi_A} = 1.$$

The probability of responding "yes" (or "no") on the selected question, if the individual does possess the attribute A, would be the same as if he or she does not. The more the Leysieffer-Warner measures of privacy protection λ_1 and λ_0 differ from unity in either direction, the more information about the characteristic under study is contained in the answer on the selected question and the lower is the personal protection against the interviewer. Once again $\lambda_1^{dir} = \lambda_0^{dir} = \infty$ (and vice versa $= 0$) applies to the direct method.

Example 3: For the Warner method with design parameter $p = 0.8$ the Leysieffer-Warner measures of privacy protection are

$$\lambda_1^W = \frac{p}{1 - p} = 4 \quad \text{and} \quad \lambda_0^W = \frac{p}{1 - p} = 4.$$

The probability of answering "yes" (or of answering "no") on the randomly selected question is 4-times higher if the respondent belongs to U_A (or to U_{A^C}) than if he belongs to U_{A^C} (or to U_A). $\triangle$

If the design parameter p of the Warner strategy was fixed optimally ($p = p_{opt}$), as it was done in Example 3, the values for λ_1^W and λ_0^W can be interpreted as the limits

$$\lambda_{1,opt}^W = \lambda_{0,opt}^W = \frac{p_{opt}}{1 - p_{opt}} \tag{9}$$

of privacy protection, which both must not be exceeded by any questioning design, if nonresponse and untruthful anwers are to be avoided and the subject as a whole is sensitive (like sexual behavior). If only the possession but not the nonpossession of attribute A is embarrassing (like drug usage), it will suffice not to exceed limit $\lambda_{1,opt}^W$.

One can think about other possible measures of privacy protection (see for example: Chaudhuri and Mukerjee (1987), p.83ff), but the principle always stays the same: There are privacy protection-limits to be kept, otherwise the respondents will

start not to respond (or not to be honest) because of the sensitivity of the character under study.

Example 4: For the Mangat-Singh version $W2*$ of $W2$ with the design parameters $p_I = 0.2$ and $p_{II} = p_{opt} = 0.8$ the Leysieffer-Warner measures of privacy protection (7) and (8) are given by

$$\lambda_1^{W2*} = \lambda_0^{W2*} = \frac{p_I + (1-p_I)\cdot p}{1 - p_I - (1-p_I)\cdot p} = \frac{0.84}{0.16} = 5.25.$$

Both limits in (9) – determined by the Warner method – are clearly exceeded. The higher efficiency of $\widehat{\pi}_A^{W2*}$ compared to $\widehat{\pi}_A^W$, that was calculated in Example 2, therefore is only seemingly and absolutely of no practical relevance. If $p_{II} = p_{opt}$, in fact only one choice for p_I is permissible, if fullresponse should occur: $p_I = 0$. $\triangle$

For the general two-stage procedure the inequalities $\lambda_1^{W2} \leq \lambda_{1,opt}^W$ and $\lambda_0^{W2} \leq \lambda_{0,opt}^W$ can only be met for all $p_I \in [0,1]$ and $p_{II} \in [0,1]$ with $min(p_{opt}, 1-p_{opt}) \leq p_I + (1-p_I)\cdot p_{II} \leq max(p_{opt}, 1-p_{opt})$. The two-stage technique under these restrictions is never more efficient than the one-stage technique, but is always more complicated in the practical application!

3 Greenberg et al.'s questioning design

The other basic idea of a randomized response technique was presented by Horvitz et al. (1967) and carried out theoretically by Greenberg et al. (1969). This questioning design (G) differs from Warner's strategy in replacing the alternative question about the membership to the subpopulation U_{A^c} by a question about the membership to a subpopulation U_B of size N_B. Elements of group U_B shall possess a completely innocuous attribute B (for instance the month of birth B or the federal state B, in which the respondent is living in), that is not related to A. Let π_B be $\frac{N_B}{N}$, the relative size of U_B and q be the design parameter of this questioning design, which – like p for Warner's method – denotes the probability of asking the question "Are you a member of group U_A?". Then

$$\hat{\pi}_A^G = \frac{\hat{\pi}_y - (1-q)\cdot \pi_B}{q} \tag{10}$$

(for $q > 0$) is an unbiased estimator of (1). For unknown π_B a modified strategy has to be used to be able to estimate π_A (Greenberg et al. (1969), p.523ff). The variance of $\hat{\pi}_A^G$ for simple random sampling without replacement is given by

$$V_{wor}(\hat{\pi}_A^G) = \frac{\pi_y\cdot(1-\pi_y)}{n\cdot q^2} - \frac{n-1}{n\cdot(N-1)}\cdot[\pi_A\cdot(1-\pi_A) + (\frac{1-q}{q})^2\cdot\pi_B\cdot(1-\pi_B)] \tag{11}$$

(Quatember and Freudenthaler (2007)) with $\pi_y = q\cdot\pi_A + (1-q)\cdot\pi_B$. For a simple random selection of sampling units with replacement (or for large populations) (11) reduces to

$$V_{wr}(\hat{\pi}_A^G) = \frac{\pi_y\cdot(1-\pi_y)}{n\cdot q^2} \tag{12}$$

(Greenberg et al. (1969), p.533).

If all other variables are fixed, the variances (11) and (12) decrease for $q \to 1$. Again exactly these values of the design parameter, which minimize the variances, perform worst in protecting the privacy of the respondents. For $q = 1$ also this strategy corresponds to the direct method with minimum privacy protection. For $q = 0$ the privacy with respect to attribute A is fully protected, but only because the question on this attribute will not be asked at all. In this case the estimation of π_A will not be possible anymore. For this method of randomized response the Leysieffer-Warner measures (7) and (8) of privacy protection are

$$\lambda_1^G = \frac{q + (1-q) \cdot \pi_B}{(1-q) \cdot \pi_B} \quad \text{and} \quad \lambda_0^G = \frac{1 - (1-q) \cdot \pi_B}{1 - q - (1-q) \cdot \pi_B}.$$

Example 5: Let the design parameter q have the same value as $p_{opt} = 0.8$ in Example 1 and $\pi_B = 0.25$. Then λ_1^G and λ_0^G are

$$\lambda_1^G = \frac{0.85}{0.05} = 17 \quad \text{and} \quad \lambda_0^G = \frac{0.95}{0.15} = 6.\dot{3}.$$

So for $q = 0.8$ both measures clearly exceed the minimum levels of privacy protection (9), that guarantee the willingness of all respondents to cooperate, which have been determined in Example 3. The respondents' answers bury more information for the interviewers on the possession or nonpossession of A than in the Warner scheme with design parameter $p = p_{opt} = 0.8$ and therefore their privacy is not kept in the same way. This means, that under our assumptions – by choosing $q = 0.8$ – fullresponse will no longer be present. The standard deviation of $\widehat{\pi}_A^G$ being $5.091 \cdot 10^{-2}$ is therefore only of theoretical interest.

To compare the efficiency of both methods at kept limits of privacy protection, for the design parameter q obviously a smaller value than 0.8 has to be chosen, if $\pi_B = 0.25$. To keep condition $\lambda_1^G \leq 4$, the maximum for the design parameter with the minimum variance of $\hat{\pi}_A^G$ is $q = q_{opt} = \frac{3}{7}$. In this case $\lambda_0^G = 2$, which means that a "no"-answer does protect the privacy more than a "yes"-answer but both limits of the respondent's protection are met. For simple random sampling the Greenberg et al. estimator (10) with design parameter $q_{opt} = \frac{3}{7}$ has a standard deviation of $9.798 \cdot 10^{-2}$. Compared under just kept restrictions of privacy protection Warner's questioning design in our example is more efficient than Greenberg et al.'s, if $\pi_B = 0.25$.

But if we can choose a nonsensitive attribute B with $\pi_B = 0.5$, the optimum design parameter will be $q_{opt} = 0.6$. With these parameters both ratios λ_1^G and λ_0^G equal 4 and the standard deviation of $\hat{\pi}_A^G$ is $7.775 \cdot 10^{-2}$. Thus at exactly the same levels of privacy protection the two methods W and G are equally efficient!

Choosing $\pi_B = 1$ leads to $q_{opt} = 0.75$, if we want to have $\lambda_1^G = 4$. But then the interviewer is able to conclude from a „no"-answer directly on the respondents' nonpossession of A, which means $\lambda_0^G = \infty$. In this case the strategy can only be used, if the possession but not the nonpossession of A would be sensitive. The standard deviation of the estimator would then be $6.532 \cdot 10^{-2}$. $\triangle$

Example 5 shows, that Greenberg et al.'s estimator (10) can be made more accurate, if only the possession of the attribute A is sensitive and therefore the privacy

of the respondent needs only to be protected for one of the two possible answers! Like the two-stage Warner questioning design a generalised two-stage version $G2$ of Greenberg et al.'s one-stage strategy G draws at the first stage the question "Do you belong to group U_A?" with probability q_I, whereas with probability $1 - q_I$ the respondent is referred to Greenberg et al.'s one-stage design with design probability q_{II}. As this strategy is obviously the same as the one-stage strategy with design parameter $q = q_I + (1 - q_I) \cdot q_{II}$ the unbiased estimator for π_A is

$$\hat{\pi}_A^{G2} = \frac{\hat{\pi}_y - (1 - q_I) \cdot (1 - q_{II}) \cdot \pi_B}{q_I + (1 - q_I) \cdot q_{II}} \tag{13}$$

(for $q_I + (1 - q_I) \cdot q_{II} > 0$). The variances of this estimator for simple random sampling without and with replacement follow from (11) and (12), if therein the probability q is replaced by $q_I + (1 - q_I) \cdot q_{II}$. But also the variances of these two- and one-stage designs may only be compared for design parameters, that keep the minimum privacy protection for the respondents to guarantee their willingness to cooperate. Therefore the probability $q_I + (1 - q_I) \cdot q_{II}$ must not come *any close* to 1. In fact the design parameters q_I and q_{II} have to be chosen in such a way, that $\lambda_1^{G2} \leq \lambda_{1,opt}^W$ (as well as $\lambda_0^{G2} \leq \lambda_{0,opt}^W$, if the whole subject is sensitive) for all q_I and $q_{II} \in [0; 1]$.

Mangat (1992) considered a special case ($G2*$) of the two-stage strategy $G2$ with $q_{II} = q$, the design parameter of Greenberg et al.'s one-stage technique (p.84):

$$\hat{\pi}_A^{G2*} = \frac{\hat{\pi}_y - (1 - q_I) \cdot (1 - q) \cdot \pi_B}{q_I + (1 - q_I) \cdot q}. \tag{14}$$

For $q_I > 0$ the estimator (14) would (theoretically) have a lower variance than Greenberg et al.'s one-stage procedure with design parameter q. But if we assume $q = q_{opt}$, then $q_I + (1 - q_I) \cdot q$ must once again not be closer to 1 than q_{opt} itself to ensure full cooperation. And then the two-stage strategy is not at all more efficient than the one-stage design!

Example 6: Let $\pi_B = 0.5$, $q_I = 0.2$ and $q = q_{opt} = 0.6$ (see Example 5). The Leysieffer-Warner measures of privacy protection for Mangat's special two-stage version of Greenberg et al.'s strategy are

$$\lambda_1^{G2*} = \frac{0.84}{0.16} = 5.25 \quad \text{and} \quad \lambda_0^{G2*} = \frac{0.84}{0.16} = 5.25.$$

As certainly λ_1^{G2*} must not exceed $\lambda_{1,opt}^W$, the probability $q_I + (1 - q_I) \cdot q = 0.68$ is too close to 1 and therefore the estimator's standard deviation of $6.713 \cdot 10^{-2}$ once again is only of theoretical interest. $\triangle$

4 Examples of other one- and two-stage strategies

In fact all one-stage techniques can be transformed into strategies with two randomization steps. To demonstrate the practical benefits of our considerations in this section we will apply them to two "two-stage" randomized response strategies consisting of three questions or statements. The first one ($S2$) consists of a first

stage, in which each individual is asked with probability r_I the question "Are you a member of group U_A?". With probability $1 - r_I$ he or she is referred to stage II, where the interviewee has to answer the same question with probability r_{II1}, just to state "yes" with probability r_{II2} or "no" with probability $1 - r_{II3}$. Like the two-stage techniques discussed in section 2 of this paper this one can be described equivalently as a (less complicated) one-stage procedure (S) of the same quality: The individual is asked the question on membership of group U_A with the total probability of $r_1 = r_I + (1 - r_I) \cdot r_{II1}$. With probability $r_2 = (1 - r_I) \cdot r_{II2}$ he or she is instructed to say "yes" and with probability $r_3 = (1 - r_I) \cdot (1 - r_{II3})$ to say "no" ($r_1 + r_2 + r_3 = 1$). Then an unbiased estimator of π_A is

$$\hat{\pi}_A^S = \frac{\hat{\pi}_y - r_2}{r_1} \tag{15}$$

($r_1 > 0$). The variance of this estimator for simple random sampling without replacement is

$$V_{wor}(\hat{\pi}_A^S) = \frac{\pi_y \cdot (1 - \pi_y)}{n \cdot r_1^2} - \frac{(n-1) \cdot \pi_A \cdot (1 - \pi_A)}{n \cdot (N-1)} \tag{16}$$

and for with replacement it is

$$V_{wr}(\hat{\pi}_A^S) = \frac{\pi_y \cdot (1 - \pi_y)}{n \cdot r_1^2}. \tag{17}$$

For the proofs of (16) and (17) see the Appendix.

For this questioning the Leysieffer-Warner measures of privacy protection (7) and (8) design are given by

$$\lambda_1^S = \frac{1 - r_3}{r_2} \quad \text{and} \quad \lambda_0^S = \frac{1 - r_2}{r_3}.$$

Example 7: To ensure that the limit $\lambda_1^S = \lambda_{1,opt}^W = 4$ can be observed, the design parameters of strategy S have to satisfy $r_1 = 3r_2 \leq \frac{3}{4}$. It shows that the minimum standard deviation of (15) under this condition is observed for $r_1 = 0.75$ and $r_2 = 0.25$. For simple random sampling in large populations this minimum is $6.532 \cdot 10^{-2}$. Because in this case $\lambda_0^S = \infty$, these values for the design parameters must only be selected, if the nonpossession of attribute A is not at all embarrassing. In this case this questioning design is exactly equally efficient as Greenberg et al.'s at the same levels of privacy protection (see: Example 6).
But if not only the membership to group U_A but also to group U_{A^c} is sensitive, under the additional condition $\lambda_0^S \leq 4$ the minimum standard deviation of $7.775 \cdot 10^{-2}$ is found for $r_1 = 0.6$ and $r_2 = r_3 = 0.2$. This means that in our example the randomized response technique S is exactly as efficient as Warner's and as Greenberg et al.'s strategy, when we compare all of them at kept limits for both measures of privacy protection! $\triangle$

A special case ($S2*$) of the two-stage version of technique S with $r_{II1} = p$ (p being the design parameter of strategy W) was presented by Singh et al. (1995). Their proof of the estimator $\hat{\pi}_A^{S2*}$ being then more efficient than the estimator $\hat{\pi}_A^W$

of Warner's technique does not make any account of the respondent's privacy protection (p.268f).

Example 8: For the questioning design $S2*$ with design parameters $r_I = 0$ and $r_{II1} = r_1 = p_{opt} = 0.8$, $r_{II2} = r_2 = 0.05$ and $r_{II3} = r_3 = 0.15$ the Leysieffer-Warner measures are given by

$$\lambda_1^{S2*} = \frac{0.85}{0.05} = 17 \quad \text{and} \quad \lambda_0^{S2*} = \frac{0.95}{0.15} = 6.\dot{3}.$$

In this case the standard deviation of (15) would be $5.091 \cdot 10^{-2}$ equalling the efficiency of strategy G at the same levels of privacy protection (Example 5). (Choosing $r_I > 0$ would increase λ_1^{S2*} and λ_0^{S2*} and reduce this standard deviation.) But as the privacy of the respondents would be badly protected this „efficiency" is only of theoretical relevance. Actually such a choice of the design parameters under our assumptions would lead to nonresponse and untruthful answers. $\triangle$

In a Warner-related randomized response technique T with three statements the interviewee with probability s_1 has to answer the question "Are you a member of group U_A?", with probability s_2 he or she is asked "Are you a member of group U_{A^c}?" and with probability s_3 the respondent is instructed to say "no" ($s_1 + s_2 + s_3 = 1$). The unbiased estimator of π_A is given by

$$\hat{\pi}_A^T = \frac{\hat{\pi}_y - s_2}{s_1 - s_2} \tag{18}$$

($s_1 \neq s_2$). The variance of this estimator for simple random sampling without replacement is

$$V_{wor}(\hat{\pi}_A^T) = \frac{\pi_A \cdot (1 - \pi_A)}{n} \cdot \frac{N - n}{N - 1} + \frac{s_1 \cdot (1 - s_1) \cdot \pi_A + s_2 \cdot (1 - s_2) \cdot (1 - \pi_A)}{n \cdot (s_1 - s_2)^2} \tag{19}$$

and for with replacement it reduces to

$$V_{wr}(\hat{\pi}_A^T) = \frac{\pi_A \cdot (1 - \pi_A)}{n} + \frac{s_1 \cdot (1 - s_1) \cdot \pi_A + s_2 \cdot (1 - s_2) \cdot (1 - \pi_A)}{n \cdot (s_1 - s_2)^2}. \tag{20}$$

For the proofs of (19) and (20) see the Appendix.

For this questioning design the Leysieffer-Warner measures of privacy protection are given by

$$\lambda_1^T = \frac{s_1}{s_2} \quad \text{and} \quad \lambda_0^T = \frac{1 - s_2}{1 - s_1}.$$

Example 9: If both the limits $\lambda_{1,opt}^W = 4$ and $\lambda_{0,opt}^W = 4$ of the minimum privacy protection should be kept, the design parameters of strategy T have to satisfy $s_1 \leq 4s_2 \leq 0.8$. The minimum variance of this questioning design is achieved, if $s_1 = 0.8$ and $s_2 = 0.2$. But in this case T reduces to W. This means, that at kept levels of privacy protection, the questioning design T cannot be more efficient than Warner's strategy. Moreover this estimator cannot be made more accurate, if only the possession of A is sensitive. $\triangle$

A two-stage version $T2$ of strategy T can be described by asking the question about the membership to U_A with probability s_I at stage I. With a probability of $1-s_I$ the interviewee is referred to a second stage, where the same question is asked with probability s_{II1}, with probability s_{II2} the question „Do you belong to group U_{A^c}?" is asked and with the remaining probability s_{II3} the respondent is instructed to state „no" ($s_{II1} + s_{II2} + s_{II3} = 1$). This is the same as the one-stage design T with the design parameters $s_1 = s_I + (1 - s_I) \cdot s_{II1}$, $s_2 = (1 - s_I) \cdot s_{II2}$ and $s_3 = (1 - s_I) \cdot s_{II3}$. When we look at the special case $T2*$ of the two-stage strategy with $s_{II1} = p$, which is the Warner design parameter, then an estimator $\widehat{\pi}_A^{T2*}$ would be more efficient than $\widehat{\pi}_A^W$, if $s_1 > max(p, 1-p)$. If p was chosen optimally with respect to the efficiency under the condition of fullresponse this would not be of any practical relevance.

Example 10: For the special case $T2*$ of $T2$ with $s_I = 0.25$, $s_{II1} = p_{opt} = 0.8$, $s_{II2} = 0.0\dot{6}$ and $s_{II3} = 0.1\dot{3}$ the measures of privacy protection are given by

$$\lambda_1^{T2*} = \frac{0.85}{0.05} = 17 \quad \text{and} \quad \lambda_0^{T2*} = \frac{0.95}{0.15} = 6.\dot{3}.$$

and therefore the standard deviation of (18) would once again only theoretically be $5.091 \cdot 10^{-2}$. $\triangle$

5 Conclusions

Randomized response strategies are an opportunity to estimate unbiasedly parameters of sensitive attributes in samples, where the direct questioning would result in nonresponse and biased estimation of unknown magnitude. Assuming that the used questioning design convinces the interviewee to cooperate, if certain limits of privacy protection are kept, it was shown in this paper for a selection of methods, that it is not only useful, but necessary, to compare the efficiency of all of these techniques only at unique optimum levels of privacy protection. Otherwise we will return to the starting point of the problem, which is characterized by nonresponse and untruthful answering.

The degree of privacy protection provided by a given questioning design with certain design parameters herein is calculated by ratios of conditional (a priori) probabilities. These measures of Leysieffer and Warner (1976) equal unity, if the privacy is fully protected by the design, which means that the probability of a certain answer is the same if the respondent does or does not belong to group U_A. Fixing limits for these measures within the (simpliest) Warner strategy, which must be kept by any questioning design, distinguishes designs that do ensure the cooperation of the sample units from others that do not.

The continuous examples 1-10 show, that the two-stage versions of the basic strategies under these considerations simply *cannot* be more efficient than the one-stage ones at the same level of privacy protection as they are nothing else than the one-stage procedures with the randomization of question (or statements) extended to two stages. Furthermore they also demonstrate, that it makes sense to distinguish also between questioning designs for subjects being sensitive as a whole (like sexual behaviour) and for subjects, where only the possession of a certain attribute is

Design k	Example	λ_1^k	λ_0^k	$[V(\widehat{\pi}_A^k)]^{1/2}$ $(\cdot 10^{-2})$
dir	1	∞	∞	4
W	1 and 3	4	4	7.775
W2*	2 and 4	5.25	5.25	6.731
G	5	17	$6.\dot{3}$	5.091
G	5	4	2	9.798
G	5	4	4	7.775
G	5	4	∞	6.532
G2*	6	5.25	5.25	6.713
S	7	4	4	7.775
S	7	4	∞	6.532
S2*	8	17	$6.\dot{3}$	5.091
T	9	4	4	7.775
T2*	10	17	$6.\dot{3}$	5.091

Table 1: The efficiency of the questioning designs of examples 1-10

embarrassing but not the nonpossession (like drug usage). All the methods compared in the examples 1-10 for large populations and certain design parameters indeed perform exactly equally well at the same levels of privacy protection.

Therefore for a serious comparison of the efficiency of randomized response strategies the kind of subject under study has to be taken into account as well as the level of privacy protection that a questioning design with certain values for the design parameters can provide.

6 Appendix

Proof of (16) and (17):
The variance of estimator (15) is given by

$$V(\hat{\pi}_A^S) = \frac{1}{r_1^2} \cdot V(\hat{\pi}_y) = \frac{1}{n^2 \cdot r_1^2} \cdot V(\sum\nolimits_s y_i). \tag{21}$$

For respondent i the variable y_i is defined as in section 2 and

$$x_i = \begin{cases} 1 & \text{if } i \text{ is asked the question on membership of } U_A, \\ 0 & \text{otherwise,} \end{cases}$$

$$v_i = \begin{cases} 1 & \text{if } i \text{ possesses the attribute } A, \\ 0 & \text{otherwise,} \end{cases}$$

$$w_i = \begin{cases} 1 & \text{if } i \text{ is instructed to say "yes",} \\ 0 & \text{otherwise,} \end{cases}$$

so that $y_i = x_i \cdot v_i + w_i$. The variance of the number of "yes"-answers in the sample is

$$V(\sum\nolimits_s y_i) = E(\sum\nolimits_s y_i^2) + E(\sum\nolimits_{s(i \neq j)} y_i \cdot y_j) - E^2(\sum\nolimits_s y_i). \tag{22}$$

For both without and with replacement simple random sampling the first summand of (22) results in

$$E(\sum\nolimits_s y_i^2) = E(\sum\nolimits_s y_i) = n \cdot (r_1 \cdot \pi_A + r_2). \tag{23}$$

The second summand of (22) is

$$E(\sum\nolimits_{s(i\neq j)} y_i \cdot y_j) = n \cdot (n-1) \cdot E(x_i \cdot x_j \cdot v_i \cdot v_j + x_i \cdot v_i \cdot w_j + w_i \cdot x_j \cdot v_j + w_i \cdot w_j).$$

Because of

$$E(v_i \cdot v_j) = \begin{cases} \frac{\pi_A \cdot (N\pi_A - 1)}{N-1} & \text{for simple random sampling without replacement,} \\ \pi_A^2 & \text{for simple random sampling with replacement} \end{cases} \tag{24}$$

for a without replacement selection of sample units we get

$$E(\sum\nolimits_{s(i\neq j)} y_i \cdot y_j) = n \cdot (n-1) \cdot (r_1^2 \cdot \frac{\pi_A \cdot (N\pi_A - 1)}{N-1} + 2 \cdot r_1 \cdot r_2 \cdot \pi_A + r_2^2) \tag{25}$$

and for a with replacement selection we have

$$E(\sum\nolimits_{s(i\neq j)} y_i \cdot y_j) = n \cdot (n-1) \cdot (r_1^2 \cdot \pi_A^2 + 2 \cdot r_1 \cdot r_2 \cdot \pi_A + r_2^2). \tag{26}$$

The subtrahend on the right side of (22) for both sampling methods is given by

$$E^2(\sum\nolimits_s y_i) = n^2 \cdot (r_1^2 \cdot \pi_A^2 + 2 \cdot r_1 \cdot r_2 \cdot \pi_A + r_2^2). \tag{27}$$

For simple random sampling without replacement the variance (21) of $\widehat{\pi}_A^S$ results in (16) by inserting (23), (25) and (27) into (22). If the sample is drawn with replacement using (26) instead of (25) results in (17). $\triangle$

Proof of (19) and (20):
The variance of estimator (18) is given by

$$V(\hat{\pi}_A^T) = \frac{1}{(s_1 - s_2)^2} \cdot V(\hat{\pi}_y) = \frac{1}{n^2 \cdot (s_1 - s_2)^2} \cdot V(\sum\nolimits_s y_i). \tag{28}$$

For respondent i the variables x_i, y_i, v_i are defined as above and

$$z_i = \begin{cases} 1 & \text{if } i \text{ is asked the question on membership of } U_{A^c}, \\ 0 & \text{otherwise.} \end{cases}$$

Then $y_i = x_i \cdot v_i + z_i - z_i \cdot v_i$. The variance of the number of "yes"-answers in the sample is given by (22). For both without and with replacement simple random sampling its first summand results in

$$E(\sum\nolimits_s y_i^2) = E(\sum\nolimits_s y_i) = n \cdot (s_1 \cdot \pi_A + s_2 - s_2 \cdot \pi_A). \tag{29}$$

The second summand of (22) is

$$\begin{aligned} E(\sum\nolimits_{s(i\neq j)} y_i \cdot y_j) &= n\cdot(n-1)\cdot E(x_i\cdot x_j\cdot v_i\cdot v_j + x_i\cdot v_i\cdot z_j - x_i\cdot v_i\cdot v_j\cdot z_j + \\ &\quad + z_i\cdot x_j\cdot v_j + z_i\cdot z_j - z_i\cdot z_j\cdot v_j - z_i\cdot v_i\cdot v_j\cdot x_j - v_i\cdot z_i\cdot z_j + \\ &\quad + z_i\cdot z_j\cdot v_i\cdot v_j). \end{aligned}$$

Because of (24) for a without replacement selection of sample units we get

$$\begin{aligned} E(\sum\nolimits_{s(i\neq j)} y_i \cdot y_j) &= n\cdot(n-1)\cdot(s_1^2\cdot\frac{\pi_A\cdot(N\pi_A-1)}{N-1} + 2\cdot s_1\cdot s_2\cdot\pi_A - \\ &\quad -2\cdot s_1\cdot s_2\cdot\frac{\pi_A\cdot(N\pi_A-1)}{N-1} + s_2^2 - 2\cdot s_2^2\cdot\pi_A + \\ &\quad + s_2^2\cdot\frac{\pi_A\cdot(N\pi_A-1)}{N-1}). \end{aligned} \tag{30}$$

And for with replacement sampling we have

$$\begin{aligned} E(\sum\nolimits_{s(i\neq j)} y_i \cdot y_j) &= n\cdot(n-1)\cdot(s_1^2\cdot\pi_A^2 + 2\cdot s_1\cdot s_2\cdot\pi_A - 2\cdot s_1\cdot s_2\cdot\pi_A^2 + s_2^2 - \\ &\quad -2\cdot s_2^2\cdot\pi_A + s_2^2\cdot\pi_A^2). \end{aligned} \tag{31}$$

For strategy T the subtrahend on the right side of (22) for both sampling methods is given by

$$E^2(\sum\nolimits_s y_i) = n^2\cdot(s_1^2\cdot\pi_A^2 + s_2^2 + s_2^2\cdot\pi_A^2 + 2\cdot s_1\cdot s_2\cdot\pi_A - 2\cdot s_1\cdot s_2\cdot\pi_A^2 - 2\cdot s_2^2\cdot\pi_A). \tag{32}$$

If the sample units are selected randomly without replacement the variance (21) results in (19) by inserting (29), (30) and (32) into (22). For the with replacement case using (31) instead of (30) gives (20) for the variance of the estimator $\widehat{\pi}_A^T$. $\triangle$

References

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

Greenberg, B., A.-L. Abul-Ela, W. Simmons, and D. Horvitz (1969). The Unrelated Question Randomized Response Model: Theoretical Framework. *Journal of the American Statistical Association 64*, 520–539.

Horvitz, D., B. Shah, and W. Simmons (1967). The unrelated question randomized response model. *1967 Social Statistics Section Proceedings of the American Statistical Association*, 65–72.

Kim, J.-I. and J. Flueck (1978). Modifications of the Randomized Response Technique for Sampling Without Replacement. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 346–350.

Leysieffer, F. and S. Warner (1976). Respondent Jeopardy and Optimal Designs in Randomized Response Models. *Journal of the American Statistical Association 71*, 649–656.

Mangat, N. (1992). Two Stage Randomized Response Sampling Procedure Using Unrelated question. *Journal of the Indian Society of Agricultural Statistics 44*, 82–87.

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

Quatember, A. and C. Freudenthaler (2007). Ein Vergleich randomisierter Antworttechniken bei Ziehen ohne Zurcklegen (in german). *Austrian Journal of Statistics*, to appear.

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

Warner, S. (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association 60*, 63–69.