New Randomized Response Technique for Estimating the Population Total of a Quantitative Variable

Jaromír Antoch[1] | *Charles University, Prague, Czech Republic*
Francesco Mola[2] | *Università di Cagliari, Cagliari, Italy*
Ondřej Vozár[3] | *Prague University of Economics and Business, Prague, Czech Republic*

Received 10.2.2022 (revision received 30.3.2022) **Accepted** (reviewed) 22.4.2022 **Published** 17.6.2022

Abstract

A new randomized response technique for estimating the population total, or the population mean of a quantitative variable is proposed. It provides a high degree of protection to the respondents because they never report their data. Therefore, it may be favorably perceived by them and increase their willingness to cooperate. Instead of revealing the true value of the characteristic under investigation, the respondent only states whether the value is greater (or smaller) than a number which is selected by him/her at random and is unknown to the interviewer. For each respondent, this number, a sort of individual threshold, is generated as a pseudorandom number. Furthermore, two modifications of the proposed technique are presented. The first modification assumes that the interviewer also knows the generated random number. The second modification deals with the issue that, for certain variables, such as income, it may be embarrassing for the respondents to report either high or low values. Thus, depending on the value of the fixed threshold (unknown to the respondent), the respondent is asked different questions to avoid being embarrassed. The suggested approach is applied in detail to the simple random sampling without replacement, but it can be, after a straightforward modification, applied to many sampling schemes, including cluster sampling, two-stage sampling, or stratified sampling. The results of the simulations illustrate the behavior of the proposed technique.

Keywords	DOI	JEL
Survey sampling, population total, Horvitz-Thompson's estimator, randomized response techniques, simple random sampling	*https://doi.org/10.54694/stat.2022.11*	*C83, J30*

[1]Charles University, Fac. of Mathematics and Physics, Sokolovská 83, CZ – 186 75 Prague 8 – Karlín, Czech Republic and Prague University of Economics and Business, Fac. of Informatics and Statistics, W. Churchill Sq. 4, CZ – 130 67 Prague 3, Czech Republic.

[2]Università di Cagliari, Fac. di Economia, viale S. Ignazio da Laconi 17, I – 09123 Cagliari, Italy.

[3]Prague University of Economics and Business, Fac. of Informatics and Statistics, W. Churchill Sq. 4, CZ – 130 67 Prague 3, Czech Republic and Czech Statistical Office, Na padesátém 3268/81, CZ – 100 82 Prague 10, Czech Republic.

Editorial note: The article was written in TEX and graphically prepared for print by the authors.

INTRODUCTION

A steady decline in response rates has been reported in many surveys in most countries around the world, see, e.g., Steeh (2001) or Stoop (2005). This decline is observed regardless of the mode of the survey, e.g., face-to-face survey, paper/electronic questionnaire, internet survey, or telephone interviewing. Furthermore, this trend has continued despite additional procedures aimed at reducing refusal and increasing contact rates; see Brick (2013) among others.

The growing concern about "invasion of privacy" therefore represents an important challenge for statisticians. Quite naturally, a respondent may be hesitant or even evasive in providing any information which may indicate a deviation from a social or legal norm and/or which he/she feels that might be used against him/her some time later. Therefore, if we ask sensitive or pertinent questions in a survey, conscious reporting of false values would often occur, see Särndal et al. (1992:547). Unfortunately, standard techniques such as reweighting or model-based imputation cannot usually be applied; for a detailed discussion see Särndal et al. (1992:547) or Särndal and Lundström (2005). On the other hand, this issue can be resolved, at least partially, using randomized response techniques (RRT). Comprehensive information on the broad scope of methods and theoretical foundations of RRT can be found in Chaudhuri (2017), Chaudhuri and Christofides (2013), Chaudhuri and Mukerjee (1988), Fox (2016) or Chaudhuri et al. (2016) among others.

For all of the reasons mentioned above, different RRTs have been developed with the goal of reducing the nonresponse rate and obtaining unbiased estimates. These techniques began with a seminal paper Warner (1965), aimed at estimating the proportion of people in a given population with sensitive characteristics, such as substance abuse, unacceptable behavior, criminal past, controversial opinions, etc. In Eriksson (1973) and in Chaudhuri (1987) the authors modified Warner's method to estimate the population total of a quantitative variable. However, in our opinion, these "standard RRTs" aimed at estimating the population total are rather complicated and demanding on both the respondents and the survey statisticians for various real-life applications; see also the discussion in Chaudhuri (2017). They require "nontrivial arithmetic operations" from respondent within the Chaudhuri's approach, while the survey statistician must expend a lot of effort related with the design of suitable randomization devices to be used for masking the sensitive variables in the Eriksson's approach.

Despite their advantages, practically all RRTs suffer from larger or smaller limitations, especially in the following.

- Lack of reproducibility.
- Lack of trust from respondents because the randomization device is controlled by the interviewer.
- Higher cost and higher variance of the estimators due to the use of random devices.

To avoid at least partially these limitations, already long time ago the statisticians suggested other approaches not requiring any random devices, These so-called *non-randomized response (NRR)* techniques are typically based on auxiliary questions, instead on random devices, and their alternative designs include, but are not limited to, unrelated question design, contamination design, multiple trials, and quantitative data design. Recently, researchers revitalized these ideas; see a series of papers by Tang, Tian, Wu, and their followers. To the best of our knowledge, they concentrated mainly on estimating proportions, not the totals. The NRR techniques are presented in detail in the monograph Tian and Tang (2014).

In any case, when suggesting any randomization device, we should always keep in mind that the main issue is not whether the in-person interviewer or telephone interviewer knows

the random numbers or the outcome of other random mechanisms used, but whether the random number is given back to the researcher evaluating the survey or to the survey sponsor. Personally, we prefer that the interviewer checks the methodology, not the realization itself.

Finally, we would like to point out that the question of credibility is not only a matter for statisticians, but more and more a task for psychologists. While statisticians must suggest procedures that are "sufficiently random" in their eyes, psychologists must find and offer ways to convince the respondents that they are not cheated. Unfortunately, a detailed discussion of this topic would go beyond the scope of this paper.

In this paper, we propose a method which is simpler in comparison with those proposed previously and which is practically applicable. The respondent is only asked whether the value of a sensitive variable reaches at least a certain random lower bound. This technique and its modifications are developed in detail and applied to simple random sampling without replacement. Their pros and cons are thoroughly discussed and illustrated using simulations.

The main advantages of the suggested method include the ease of implementation, simple use by the respondent, and practically acceptable precision. Moreover, respondents' privacy is well protected because they never report the true value of the sensitive variable. Unlike in Chaudhuri's or Eriksson's approach, there is no issue with the physical random device design. A disadvantage may be, from a certain point of view, a lower degree of confidence in anonymity due to the extrinsic device/technique used for generating random numbers.

The paper is organized as follows. In Section 1, selected issues of the RRTs for the estimation of the population total, or population mean, are concisely discussed. In Section 2, a new randomized response technique and its two modifications are proposed, their properties studied and the goals for future work summarized. Section 3 illustrates the suggested ideas with the aid of a simulation study. The main conclusions of the paper follow.

1 SELECTED REMARKS ON RRT INTENDED TO ESTIMATE POPULATION TOTAL AND THEIR PROPERTIES

Consider a finite population $U = \{1, \ldots, N\}$ of N identifiable units, where each unit can be unambiguously identified by its label. Let Y be a sensitive quantitative variable. The objective of the survey is to estimate the population total $t_Y = \sum_{i \in U} Y_i$ or, alternatively, the population mean $\overline{t}_Y = t_Y/N$, of the variable surveyed. To do this, we use a random sample s selected with probability $p(s)$, described by a sampling plan with a fixed sample size n. Let us denote by π_i the probability of inclusion of the i^{th} element in the sample, that is, $\pi_i = \sum_{s \ni i} p(s)$, and by ξ_i the indicator of inclusion of the i^{th} element in the sample s, i.e., $\xi_i = 1$ if $s \ni i$ and $\xi_i = 0$ otherwise. We do not introduce all notions from scratch and refer the reader to Särndal et al. (1992:547) or the more rigorous monograph Tillé (2006).

As argued above, in practice it is often impossible to obtain the values of the surveyed variable Y in sufficient quality due to its sensitivity. Therefore, statisticians try to obtain from each respondent at least a randomized response Z that is correlated to Y. This randomization of the responses must be carried out independently for each population unit in the sample and independently of the sampling plan $p(s)$.

In such a case, the survey has two phases. First, a sample s is selected from U and then, given s, responses Z_i are realized using the selected RRT. We denote the corresponding probability distributions by $p(s)$ and $q\big(r\,|s\big)$. In this setting, the notions of expected value, unbiasedness, and variance are tied to a two-fold averaging process.

- Over all possible samples s that can be drawn using the selected sampling plan $p(s)$.
- Over all possible response sets r that can be realized given s under the response distribution $q\big(r\,|s\big)$.

In the sequel, we follow the literature and, where appropriate, denote the expectation operators with respect to these two distributions by E_p and E_q, respectively.

In a direct survey, the population total t_Y is usually estimated from the observed values Y_i using a linear estimator $t_s = \sum_{i \in s} b_{si} Y_i$, where the weights b_{si} follow the unbiasedness constraint $\sum_{s \ni i} p(s) b_{si} = 1,\ i = 1, \ldots, N$. If $\pi_i > 0\ \forall i \in U$, then Horvitz-Thompson's estimator

$$t_s^{HT} = \sum_{i \in s} \frac{Y_i}{\pi_i} \tag{1}$$

is a linear unbiased estimator with weights $b_{si} = 1/\pi_i$, and $\mathrm{E}_p\left(t_s^{HT}\right) = t_Y$, see Horvitz and Thompson (1952) or Section 2.8 in Tillé (2006) for details.

If the survey is conducted using RRT, the true values of Y_i for the sample s are unknown and, instead of them, the values of random variables Z_i correlated to Y_i are collected. Variables Z_i are usually further transformed into another variables R_i, which are more suitable for the construction of the desired estimator, and then the population total is typically estimated using a Horvitz-Thompson's type estimator

$$t_s^{HT,R} = \sum_{i \in s} \frac{R_i}{\pi_i}. \tag{2}$$

Suppose now that we have an estimator (a formula, or a computational procedure) for estimating the population total t_Y or population mean $\overline{t}_Y$; we denote it by t_Y^R and $\overline{t}_Y^R$, respectively. The subscript R emphasizes that the estimator is based on the values of R_i, i.e., on randomized responses. Furthermore, we assume that the randomized responses R_i follow a model for which it holds $\mathrm{E}\left(R_i\right) = Y_i$, $\mathrm{Var}\left(R_i\right) = \phi_i\ \forall i \in U$, and $\mathrm{Cov}\left(R_i, R_j\right) = 0\ \forall i \neq j,\ i, j \in U$. Note that the variance function ϕ_i of a randomized response R_i is a function of Y_i.

Recall that the estimator t_Y^R of the population total t_Y *is conditionally unbiased*, if the conditional expectation of t_Y^R given the sample s is equal to the current estimator t_s that would be obtained if no randomization took place, that is, if $\mathrm{E}_q\left(t_Y^R \mid s\right) = t_s$. The subscript s indicates that the "usual" estimator based on the nonrandomized sample, for example the Horvitz-Thompson's one, is used, and $\mathrm{E}_q\left(t_Y^R \,\middle|\, s\right)$ stands for the conditional expectation of t_Y^R given the sample s with respect to the distribution induced by the randomization of responses. For the estimator $\overline{t}_Y^R$ of the population mean, we proceed analogously.

If t_Y^R is conditionally unbiased and t_s is unbiased, then t_Y^R is also unbiased, since $\mathrm{E}\left(t_Y^R\right) = \mathrm{E}_p\big(\mathrm{E}_q(t_Y^R \mid s)\big) = \mathrm{E}_p\big(t_s\big) = t_Y$. Analogously, it holds $\mathrm{E}\left(\overline{t}_Y^R\right) = \overline{t}_Y$. Moreover, by a standard formula of the probability theory, we get the variance of t_Y^R in the form

$$\begin{aligned} \mathrm{Var}\left(t_Y^R\right) &= \mathrm{E}_p\Big(\mathrm{Var}_q\big(t_Y^R \,\big|\, s\big)\Big) + \mathrm{Var}_p\Big(\mathrm{E}_q\big(t_Y^R \,\big|\, s\big)\Big) \\ &= \mathrm{E}_p\Big(\mathrm{Var}_q\big(t_Y^R \,\big|\, s\big)\Big) + \mathrm{Var}_p\big(t_s\big). \end{aligned} \tag{3}$$

The second term on the right-hand side of (3) is, obviously, the variance of the estimator that would apply if no randomization of responses was deemed necessary, while the first term represents the increase of the variance produced by the randomization. In other words, the two terms on the right-hand side of (3) represent, respectively, *contribution by randomized response technique used* and *contribution by sampling variation* to the total variance of t_Y^R. When treating $\overline{t}_Y^R$, we proceed analogously.

Because the design-based expression for the variance of t_s is known for the most common sampling procedures, we can focus on the contribution of randomization and study it in more detail. For example, for the estimator $t_s^{HT,R}$ given by (2), we have

$$\begin{aligned}\mathrm{Var}\left(t_s^{HT,R}\right) &= \mathrm{E}_p\Big(\mathrm{Var}_q\big(t_s^{HT,R}\,\big|\,s\big)\Big) + \mathrm{Var}_p\Big(\mathrm{E}_q\big(t_s^{HT,R}\,\big|\,s\big)\Big)\\ &= \mathrm{E}_p\Big(\sum_{i\in U}\frac{\phi_i\xi_i}{\pi_i^2}\Big) + \mathrm{Var}_p\big(t_s^{HT}\big) = \sum_{i\in U}\frac{\phi_i}{\pi_i} + \mathrm{Var}\left(t_s^{HT}\right). \end{aligned} \tag{4}$$

Several techniques for estimating the population total were suggested in the literature. The papers Eriksson (1973) and Chaudhuri (1987) were at the origin, and became a benchmark for many following approaches. Both techniques have been further developed and improved by other researchers; see, e.g., interesting papers Arnab (1995, 1998) or Gjestvanga and Singh (2009). The ideas and a representative review of further research are presented in a monograph Chaudhuri (2017). Another type of randomization technique was suggested in a series of papers by Dalenius and his colleagues, e.g., Bourke and Dalenius (1976) or Dalenius and Vitale (1979). Among recent papers on the topic of sensitive questions in population surveys, we would like to mention, for example, papers by Kirchner (2015) and Trappmann (2014). In both of them, long lists of relevant references can be found. Finally, recall that probably the most comprehensive account of developments in sample survey theory and practice can be found in Pfeffermann and Rao (2009a,b), or in the more recent monographs Arnab (2017), Tian and Tang (2014), Tillé (2020) or Wu and Thompson (2020).

2 NEW RANDOMIZED RESPONSE TECHNIQUE

In this section, we suggest a completely different approach. Assume that the studied sensitive variable Y is non-negative and bounded from above, i.e., $0 \le Y \le M$. First, let us assume the upper bound M of the variable Y is known. Each respondent performs, independently of the others, a random experiment generating a pseudorandom number Υ from the uniform distribution on interval $(0, M)$, while the interviewer does not know this value. The respondent can generate the pseudorandom number Υ using, for example, a laptop online/offline application; for some other possibilities, see Section 2.4. The respondent then answers a simple question: "*Is the value of Y greater than Υ?*" (e.g.: "*Is your monthly income greater than Υ?*").

For certain sensitive variables, such as the total amount of alcohol consumed within a certain period, it is better to use a question: "*Is the value of Y lower than Υ?*" In such a case we recode the response $Z_{i,(0,M)}$ to $Z^{\star}_{i,(0,M)} = 1 - Z_{i,(0,M)}$, and apply the suggested RRT to $Z^{\star}_{i,(0,M)}$.

The response of the i^{th} respondent follows the alternative distribution with the parameter Y_i/M, that is

$$Z_{i,(0,M)} = \begin{cases} 1 \text{ with probability } \frac{Y_i}{M}, & \text{if } \Upsilon_i < Y_i,\\ 0 \text{ with probability } 1 - \frac{Y_i}{M}, & \text{otherwise.}\end{cases} \tag{5}$$

Therefore, $\mathrm{E}\left(Z_{i,(0,M)}\right) = P\big(\Upsilon_i < Y_i\big) = Y_i/M$ and $\mathrm{Var}\left(Z_{i,(0,M)}\right) = \big(Y_i/M\big)\big(1 - Y_i/M\big)$. Therefore, we transform $Z_{i,(0,M)}$ to $R_{i,(0,M)} = MZ_{i,(0,M)}$, for which we have

$$\mathrm{E}\left(R_{i,(0,M)}\right) = Y_i \quad \text{and} \quad \mathrm{Var}\left(R_{i,(0,M)}\right) = Y_i\big(M - Y_i\big). \tag{6}$$

2.1 Application to the simple random sampling

Consider now the situation in which the sampling plan $p(s)$ is a simple random sampling without replacement with a fixed sample size n. Denote, as in Section 1, by $\overline{t}_Y = \frac{1}{N}\sum_{i\in U} Y_i$ the population mean, by $S_Y^2 = \frac{1}{N-1}\sum_{i\in U}\left(Y_i - \overline{t}_Y\right)^2$ the population variance. In this case, the inclusion probabilities are constant, that is, $\pi_i = P(\xi_i = 1) = n/N \ \forall i \in U$.

Let the population total t_Y be estimated using the Horvitz-Thompson's type estimator

$$t_{(0,M)}^{HT,R} = \frac{N}{n}\sum_{i\in s} R_{i,(0,M)}. \tag{7}$$

It follows from (6) that this estimator is unbiased, so let us calculate its variance. For this purpose (4) can be used effectively. First, note that in the case considered $\pi_i = n/N$, and due to (6) $\phi_i = Y_i(M - Y_i)$. Second, taking into account variance of the simple random sampling without replacement, see Section 4.4 in Tillé (2006) for details, we obtain after a straightforward calculation

$$\mathrm{Var}\left(t_{(0,M)}^{HT,R}\right) = \frac{N^2}{n}\Big(\overline{t}_Y(M - \overline{t}_Y) - \frac{N-1}{N}S_Y^2\Big). \tag{8}$$

To characterize the variance of the suggested estimators more deeply and to get a more transparent understanding of the variance of the suggested RRT, we introduce two auxiliary characteristics termed *measures of concentration*. More precisely, let us denote

$$\Gamma_{Y,M} = \frac{1}{N}\sum_{i\in U}\frac{Y_i}{M}\Big(1 - \frac{Y_i}{M}\Big) = \underbrace{\frac{1}{MN}\sum_{i\in U} Y_i}_{\frac{1}{M}\overline{t}_Y} - \underbrace{\frac{1}{M^2N}\sum_{i\in U} Y_i^2}_{\frac{1}{M^2}\overline{Y^2}} = \frac{\overline{t}_Y}{M} - \frac{\overline{Y^2}}{M^2} \tag{9}$$

and

$$\Gamma_{\overline{Y},M} = \frac{\overline{t}_Y}{M}\frac{(M - \overline{t}_Y)}{M} = \frac{\overline{t}_Y}{M} - \frac{\overline{t}_Y^2}{M^2}. \tag{10}$$

In the sequel, we call $\Gamma_{Y,M}$ *the mean relative concentration measure*, and $\Gamma_{\overline{Y},M}$ *the proximity measure of the population mean* $\overline{t}_Y$ *to* $M/2$.

If Y_i are iid random variables with finite variance σ^2 and an expectation μ, then, by the law of large numbers, both $\Gamma_{Y,M}$ and $\Gamma_{\overline{Y},M}$ converge, as $N \to \infty$, with probability 1 to

$$\Gamma_{Y,M,as} = \frac{\mu}{M}\Big(1 - \frac{\mu}{M}\Big) - \frac{\sigma^2}{M^2} \quad \text{and} \quad \Gamma_{\overline{Y},M,as} = \frac{\mu}{M}\Big(1 - \frac{\mu}{M}\Big). \tag{11}$$

We call $\Gamma_{Y,M,as}$ *the asymptotic mean relative concentration measure*, and $\Gamma_{\overline{Y},M,as}$ *the asymptotic proximity measure of the population mean* $\overline{t}_Y$ *to* $\frac{M}{2}$. Note that both $\Gamma_{Y,M,as}$ and $\Gamma_{\overline{Y},M,as}$ exist if $0 \le Y_i \le M \ \forall i \in U$.

Let us focus on $\Gamma_{Y,M}$ and $\Gamma_{\overline{Y},M}$ in more detail. First, note that in our setting both are population characteristics, not random variables. Second, both take their values in the interval $[0, 1/4]$, and are equal to zero only in pathological cases when either $Y_i = 0 \ \forall i \in U$ or $Y_i = M \ \forall i \in U$. The higher these measures, the higher the variance of $t_{(0,M)}^{HT,R}$. The mean relative concentration measure $\Gamma_{Y,M}$ reaches its maximum $1/4$ when all values are at the center of the interval $(0, M)$, that is, if $Y_i = M/2 \ \forall i \in U$. The proximity measure $\Gamma_{\overline{Y},M}$ of the population mean to the center of the interval $(0, M)$ reaches its maximum $1/4$ only if the population mean is at the center of the interval, that is, $\overline{t}_Y = M/2$. This case occurs, e.g.,

when random variable Y is symmetric around the center of the interval $M/2$; this feature is certainly true for the uniform distribution on $(0, M)$.

For a fixed value of the upper bound M, population size N and sample size n, the contribution of the suggested RRT to the variance of $t^{HT,R}_{(0,M)}$ depends, up to a multiplicative constant, on $\Gamma_{Y,M}$, because it holds

$$\mathrm{E}_p\Big(\mathrm{Var}_q\big(t^{HT,R}_{(0,M)} \mid s\big)\Big) = \frac{M^2N^2}{n} \underbrace{\frac{1}{N}\sum_{i\in U}\frac{Y_i}{M}\Big(\frac{M-Y_i}{M}\Big)}_{\Gamma_{Y,M}} = \frac{M^2N^2}{n}\Gamma_{Y,M}. \tag{12}$$

Analogously, this contribution can also be expressed, up to multiplicative constants, by $\Gamma_{\overline{Y},M}$ and S^2_Y, because it holds

$$\mathrm{E}_p\Big(\mathrm{Var}\,\big(t^{HT,R}_{(0,M)} \mid s\big)\Big) = \frac{M^2N^2}{n}\Gamma_{\overline{Y},M} - \frac{N(N-1)}{n}S^2_Y. \tag{13}$$

Thus, both $\Gamma_{Y,M}$ and $\Gamma_{\overline{Y},M}$ can help us explain how the suggested RRT increases the variance of the estimator of the population total t_Y for distributions symmetrical around $M/2$, for distributions concentrated close to the center of $(0, M)$, symmetrical around $M/2$, or uniformly distributed. Moreover, they show that the suggested approach is especially suitable for skewed distributions, provided that they are concentrated around their mean values. Let us sum up: both measures of concentration help us not only to describe the variance of the estimator used, compare (12) and (13), but also to interpret it better.

Remark 1. If the values of Y are bounded both from below and above, that is, $0 < m \leq Y \leq M$, then variance of $t^{HT,R}_{(0,M)}$ can be significantly reduced by generating pseudorandom numbers Υ_i from the uniform distribution on the interval (m, M) instead on $(0, M)$. In fact, if this is the case, we replace $Z_{i,(0,M)}$, described by (5), with

$$Z_{i,(m,M)} = \begin{cases} 1 \text{ with probability } \frac{Y_i-m}{M-m}, & m \leq \Upsilon_i < Y_i, \\ 0 \text{ with probability } 1-\frac{Y_i-m}{M-m}, & \text{otherwise,} \end{cases}$$

transform these variables to $R_{i,(m,M)} = m + (M-m)Z_{i,(m,M)}$, and estimate population total t_Y analogously to (7) using the Horvitz-Thompson's type estimator

$$t^{HT,R}_{(m,M)} = \frac{N}{n}\sum_{i\in s} R_{i,(m,M)}. \tag{14}$$

It is easy to show that the variance of $t^{HT,R}_{(m,M)}$ is smaller than that of $t^{HT,R}_{(0,M)}$, that is, by the value $\frac{N^2m}{n}\big(M-\overline{t}_Y\big)$.

The immediate question arises of what happens if the interval $[m, M]$ is not set correctly. Evidently, if some values of Y_i are outside the interval $[m, M]$, then with probability 1 it holds $Z_{i,(m,M)} = 0$ if $Y_i < m$ and $Z_{i,(m,M)} = 1$ if $Y_i > M$. The bias of the suggested estimator is equal to

$$\sum_{i\in U \mid Y_i<m}\big(Y_i - m\big) + \sum_{i\in U \mid Y_i>M}\big(Y_i - M\big). \tag{15}$$

In practice, the bounds of the variable Y are often unknown. When choosing parameters m and M, a researcher should carefully consider the trade-off between bias and privacy.

While lower bound m affects mostly bias and is not very crucial to the privacy of respondents, the choice of M affects both bias and privacy. Moreover, there is also a trade-off between bias and variance of estimates; see the results of the simulations in Tables 2 – 4 in the Annex. Therefore, a reasonable guess about the empirical quantiles of the characteristics studied is vital for setting the values of m and M properly.

Let us discuss some advantages and disadvantages of our approach compared to the other techniques suggested in the literature.

- It is simple; this fact increases respondents' confidence and cooperation, and thus reduces the estimation error.
- Respondents' privacy is well protected, because they never report the true value of the sensitive variable.
- It avoids the demanding task of designing a randomization device intended for masking the surveyed variable.
- It enables to estimate the population total at an acceptable level of accuracy, see Section 3 for details. Of course, what level is acceptable depends on the survey and selected precision requirements. According to our simulations, standard errors of the estimators described up to now are at most two times higher than those of HT-estimators, see Tables 2 – 4.
- On the other hand, due to the need of a device/technique for generation random numbers, some respondents may feel a lower degree of confidence in preserving their anonymity.

Finally, we find rather problematic any comparison of our approach with other methods because their performance strongly depends on the choice of the randomizing device used. In our opinion, it is tricky to design, e.g., a deck of cards for a continuous variable with a high range, such as the income in the Czech Republic, and a reliable estimator of this type with an acceptably small variance value would need an excessively large size.

2.2 Estimators using knowledge of Υ

A natural question arises as to whether we could improve the accuracy of the suggested method. Thus, in what follows, we discuss the two modifications of the RRTs suggested in Section 2.1 and their properties in the following subsections. The heuristics behind this approach are based on the following observations. All techniques presented up to now have assumed that the interviewer does not know the outcome of the randomization device leading to the randomized response, such as the card drawn, the value of the pseudorandom number, etc. It is plausible to ask what would happen if we also knew the outcome of that random experiment on the one hand, while protecting respondents' privacy on the other one. More precisely: *Can we modify the estimator and to increase its accuracy, that is, to decrease its variance, if we also know the values of the generated pseudorandom number*? We surmise that it is feasible and suggest one possible way of reaching this goal. However, we point out that the success of the suggested approach, to a considerable extent, depends on the statistician's insight into the problem.

Assume again that the studied sensitive variable Y is non-negative and bounded from above, that is, $0 \leq Y \leq M$. Each respondent carries out, independently of the others, a random experiment generating a pseudorandom number Υ from the uniform distribution on interval $(0, M)$, and *informs the interviewer of both its value* and *whether* $\Upsilon < Y$ *or not*. For example, the response is that the simulated number has been xxx (let say 45 000 CZK) and the respondent earns more/less. To distinguish from the situation described in Section 2.1, we further assume that the corresponding random response is now described

using a dichotomous random variable

$$Z_{i,\alpha,(0,M)} = \begin{cases} 1-\alpha+2\alpha\frac{\Upsilon_i}{M}, & \text{if } \Upsilon_i < Y_i, \\ -\alpha+2\alpha\frac{\Upsilon_i}{M}, & \text{otherwise,} \end{cases} \qquad 0 \le \alpha < 1,\ i = 1,\dots,n, \tag{16}$$

where α is a tuning parameter. Its value is a priori set by the interviewer, is fixed and unknown to the respondent. This proposal is a linear combination of our initial proposal $Z_{i,(0,M)}$ given by (5) and $2\Upsilon_i/M$. Higher the value α, more weight is put on the term using the pseudorandom number Υ. For $\alpha = 0$ we have the initial method described in Section 2.1. The rule for an optimal choice of α is given later in this section.

The response of the respondent to $Z_{i,\alpha,(0,M)}$ is transformed not by the respondent, but by the interviewer off-line. The discussion about the choice of α is postponed here and will be done later.

Since $P\big(Z_{i,\alpha,(0,M)} = 1-\alpha+2\alpha\frac{\Upsilon_i}{M}\big) = P\big(\Upsilon_i < Y_i\big)$, we have

$$\mathrm{E}\left(Z_{i,\alpha,(0,M)}\right) = \frac{1}{M}\int_0^{Y_i}\Big(1-\alpha+2\alpha\frac{u}{M}\Big)\,du + \frac{1}{M}\int_{Y_i}^{M}\Big(-\alpha+2\alpha\frac{u}{M}\Big)\,du = \frac{Y_i}{M},$$

$$\mathrm{Var}\left(Z_{i,\alpha,(0,M)}\right) = \frac{1-2\alpha}{M^2}Y_i\big(M-Y_i\big) + \frac{\alpha^2}{3}.$$

Therefore, the random responses $Z_{i,\alpha,(0,M)}$ are further transformed to $R_{i,\alpha,(0,M)} = MZ_{i,\alpha,(0,M)}$, and the desired estimator of the population total t_Y is constructed analogously to (7) and (14). More precisely, we suggest using again the Horvitz-Thompson's type of estimator in the form

$$t^{HT,R}_{\alpha,(0,M)} = \frac{N}{n}\sum_{i\in s} R_{i,\alpha,(0,M)}. \tag{17}$$

It is evident that $\mathrm{E}\left(R_{i,\alpha,(0,M)}\right) = Y_i$, so the estimator (17) is unbiased. Moreover, the contribution of randomization to its variance is

$$\mathrm{E}_p\Big(\mathrm{Var}_q\big(t^{HT,R}_{\alpha,(0,M)}\,\big|\,s\big)\Big) = \frac{M^2N^2}{n}\sum_{i\in U}\Big[\,\frac{1}{N}\big(1-2\alpha\big)\frac{Y_i}{M}\Big(1-\frac{Y_i}{M}\Big) + \frac{\alpha^2}{3N}\,\Big]. \tag{18}$$

An easy calculation shows that (18) has a global minimum at $\alpha = 3\Gamma_{Y,M} \in [0, 3/4]$. If we set $\alpha_{opt} = 3\Gamma_{Y,M}$ and substitute it back to (18), then the contribution of randomization to the variance of (17) for this choice of α is

$$\begin{aligned}\mathrm{E}_p\Big(\mathrm{Var}_q\big(t^{HT,R}_{\alpha_{opt},(0,M)}\,\big|\,s\big)\Big) &= \frac{M^2N^2}{n}\sum_{i\in U}\Big[\,\big(1-6\Gamma_{Y,M}\big)\frac{1}{N}\frac{Y_i}{M}\Big(1-\frac{Y_i}{M}\Big) + \frac{3\Gamma^2_{Y,M}}{N}\,\Big] \\ &= \frac{M^2N^2}{n}\Gamma_{Y,M}\big(1-3\Gamma_{Y,M}\big).\end{aligned} \tag{19}$$

If we compare (19) with (12), we see that the knowledge of pseudorandom numbers Υ_i and the use of α_{opt} considerably decrease the variability, of course, depending on the suggested RRT. It is worth highlighting that our simulations summarized in Section 3 confirm these findings.

The conclusion that the knowledge of Υ leads to a smaller variance of the estimator is expected; see above. The reason is clear and is based on the well-known inverse relationship

that exists between the disclosure of personal information and the efficiency of estimates, that is, *the more the privacy is jeopardized the lower the variance*. For a discussion, see Chaudhuri and Mukerjee (1988), among others. In our case, the assumption that the interviewer knows Υ means that the privacy of the respondent is less protected and, consequently, we get better estimates.

Parameter α should be set to its optimal value $\alpha_{opt} = 3\Gamma_{Y,M}$, where the mean relative concentration measure $\Gamma_{Y,M}$ is introduced in Section 2, Formula (9). If the interviewer has some prior information about the mean μ and variance σ^2 values for the theoretical distribution of the surveyed variable Y, he/she should rather apply the asymptotic concentration measure (11), which can be estimated using a plug-in moment estimator. More precisely, the population mean $\overline{t}_Y$ should be replaced by μ, and the population variance S_Y^2 by σ^2. Since the population second moment $\overline{Y^2}$ can be expressed as $\frac{N-1}{N}S_Y^2 + \overline{t}_Y^2$, it is sufficient to substitute μ and σ^2 into this expression. Moreover, recall that the prior information is often available for regular surveys in official statistics, such as EU-SILC, because in such a case we can either use results from previous years updated by inflation, or we can rely on the expert opinion. If no prior information is available, we recommend choosing small values of α, such as 0.5.

Notice that if a nonnegative surveyed random variable Y is bounded not only from above but also from below, that is, $0 < m \leq Y \leq M$, we generate Υ_i from the uniform distribution on the interval (m, M), modify $Z_{i,\alpha,(0,M)}$ given by (16) to

$$Z_{i,\alpha,(m,M)} = \begin{cases} 1 - \alpha + 2\alpha\frac{\Upsilon_i - m}{M-m}, & \text{if } \Upsilon_i < Y_i, \\ -\alpha + 2\alpha\frac{\Upsilon_i - m}{M-m}, & \text{otherwise,} \end{cases} \qquad 0 \leq \alpha < 1,$$

transform $Z_{i,\alpha,(m,M)}$ to $R_{i,\alpha,(m,M)} = m + (M-m)Z_{i,\alpha,(m,M)}$, and form an estimator of the population total t_Y of the Horvitz-Thompson's type, parallel to (17), as

$$t^{HT,R}_{\alpha,(m,M)} = \frac{N}{n}\sum_{i\in s} R_{i,\alpha,(m,M)}. \tag{20}$$

Using analogous arguments as above, it is straightforward to show that $\mathrm{E}\left(R_{i,\alpha,(m,M)}\right) = Y_i$, so that the estimate (20) is again unbiased regardless of the value of the parameter α.

We must firmly emphasize that neither the information about the value of pseudorandom number Υ nor the value α enables us to guess the exact value of the sensitive variable Y, except for the case $Y = M$. In other words, knowing them does not intrude on the respondent's privacy.

The heuristics behind the proposed modification are the following:

- If the response is *YES*, then a high value of the pseudorandom number Υ implies a high value of the studied variable Y, because $Y > \Upsilon$, and these observations "considerably" increase the value of the estimator.
- However, if the response is *NO*, then a low value of the pseudorandom number Υ implies a low value of Y, because $Y \leq \Upsilon$, and these observations "considerably" decrease the value of the estimator.

Unfortunately, in both situations, that is, when the value of the response is either (too) low or (too) high, the respondent may be more prone to fabricate his/her response.

As we can see, $Z_{i,\alpha,(m,M)}$ can occasionally attain negative values, which is an obvious drawback. On the other hand, using a guess about the distribution of Y, it is possible to estimate (at least roughly) the probability of such an event. For illustration, in the case of

practical application of our approach described in Section 3, the probability of obtaining a negative $Z_{i,\alpha,(m,M)}$ is of the order 10^{-5}, and during our extensive simulations, we never met such a case. In this paper, we do not study the effect on bias and variance when setting negative values to zero.

2.3 Estimator using switching questions

We emphasize that for some characteristics, such as the monthly income of a household or alcohol consumption, it can be sensitive for respondents to report either high or low values. This led us to modify the suggested RRT approach in the following way.

Assume that a nonnegative surveyed random variable Y is bounded both from below and above, that is, $0 < m \leq Y \leq M$. First, we set a proper fixed threshold T, $m < T < M$, unknown to the respondent. Second, we generate Υ from the uniform distribution on (m, M) and, depending on whether the pseudorandom number Υ does or does not exceed this fixed threshold T, we ask one of the following questions:

i. If $\Upsilon \leq T$: "*Is the value of* Y *greater than* Υ*?*",
ii. If $\Upsilon > T$: "*Is the value of* Y *smaller or equal than* Υ*?*".

Third, for the ith respondent, we form a random variable

$$Z_{i,T,(m,M)} = \begin{cases} 1, & \text{if } \Upsilon_i \leq T \ \& \ \Upsilon_i \leq Y_i, \\ 0, & \text{if } \Upsilon_i \leq T \ \& \ \Upsilon_i > Y_i \quad \text{or} \quad \Upsilon_i > T \ \& \ \Upsilon_i \leq Y_i, \\ -1, & \text{if } \Upsilon_i > T \ \& \ \Upsilon_i > Y_i. \end{cases}$$

If we know both the response concerning the value of Y and the question asked, that is whether $\Upsilon_i \leq T$ or not, then it is easy to show that $\mathrm{E}\left(Z_{i,T,(m,M)}\right) = (T+Y_i-m-M)/(M-m)$. This advises to transform $Z_{i,T,(m,M)}$ to $R_{i,T,(m,M)} = (M-m)Z_{i,T,(m,M)}+m+M-T$, because then $\mathrm{E}\left(R_{i,T,(m,M)}\right) = Y_i$. Thus, the Horvitz-Thompson's type estimator of the population total t_Y of the form

$$t^{HT,R}_{T,(m,M)} = \frac{N}{n} \sum_{i \in s} R_{i,T,(m,M)}. \tag{21}$$

is evidently also unbiased.

As concern variance of $R_{i,T,(m,M)}$, we must distinguish between $Y_i > T$ and the complementary inequality. After a bit of tedious calculation we get, as expected, that it is always higher than the variance of $R_{i,(m,M)}$. Worse still is the fact that negative values of $Z_{i,T}$ may occur quite frequently, leading to negative values of the corresponding $R_{i,T,(m,M)}$. Looking at the results of our simulations, we observe that $t^{HT,R}_{T,(m,M)}$ can return inadmissibly low or even negative values, which is a major drawback. Moreover, we cannot find the way how to set optimal value of the threshold T minimizing $\mathrm{Var}\left(R_{i,T,(m,M)}\right)$, being another drawback.

An unbiased estimator of the population mean $\overline{t}_Y$ can be constructed in parallel. On the other hand, if we know only the response concerning the value of Y but not the question asked, in this case it is not possible to construct an estimator of the population total t_Y, respectively of the population mean $\overline{t}_Y$.

Thus, the seemingly appealing idea described in this section seems to be interesting from a theoretical point of view. We cannot recommend it for practical use automatically without prior information on the population studied, which is also illustrated by the simulations presented in Section 3.

2.4 Random number generation

In all RRTs, the choice of randomization device is probably the trickiest point. If we assume direct face-to-face interviewing, the following points describe several possibilities that might be used in our approach.

- We allow the respondent to select the random number according to some standard, e.g. the European ISO 28640:2010(en) Standard ISO. We are convinced that the existence of a standard can increase the credibility of the survey and willingness of respondents to respond truthfully. The selected random number is then used according to the RRT used.
- To those respondents who feel like "experts in the field of randomness", the reviewer can offer them the option to select a random number from the uniform distribution using their own method.
- Another possibility is, for example, to use a large deck of cards, but it would require additional calculations to find the bias of such an approach.

2.5 Open problems

There are several relevant related research issues, not treated in this paper due to its current length. We aim to concentrate on them in subsequent papers. They include, but are not limited to, the following points:

- To generalize suggested estimators to more complex sampling plans as cluster sampling, two-stage sampling, stratified sampling, etc. Moreover, it has been repeatedly emphasized during the discussions, e.g. after the presentation of our results, that the median and other quantiles are important statistics for many applications, sometimes even more important than the mean. Similarly, the question has been raised whether parallel methodology could be used in any type of regression analysis combining it, e.g., with ideas from Antoch and Janssen (1989), Pfeffermann and Rao (2009a,b) or Tillé (2020).
- To modify, where appropriate, suggested estimators to the case when pseudorandom numbers are generated not from the uniform distribution, but from the distribution that mimics the surveyed variable Y. To prepare a numerical study illustrating the effect of the distribution from which we simulate random numbers on the on the possible improvements in the performance of estimators. In the case of income covered in our simulation example, the log-logistic or log-normal distribution might be used.
- To study more profoundly effects of tuning parameters on the bias, variance, and privacy jeopardy, as well as the trade-off among the parameters and pseudorandom numbers generated from different distributions. To suggest rules of thumb for the choice of parameters m, M and α and to study optimal choice of parameters with respect to the minimization of the mean square error.
- To derive unbiased estimators of variance and to study the impact on the corresponding confidence intervals. To study the effect on bias and variance of the suggested procedures when treating possible negative values as zeros.
- To compare our proposal with that of the unrelated question model suggested originally in Greenberg (1971).
- To find approximate formulae for sample sizes with required margin of error.

3 SIMULATION STUDY

In many countries, income is recognized as private and (highly) sensitive information. Respondents often refuse to respond at all or provide strongly biased responses. This in particular happens if their income is (very) high or (very) low. This leads us to assess the

performance of the proposed RRTs through a simulation study using Czech wage data from the Average Earnings Information System (IPSV) of the Ministry of Labor and Social Affairs of the Czech Republic.

Based on the extensive analysis of monthly wage statistics provided by IPSV for the years 2004 – 2014, Vrabec and Marek (2016) recommended to model wages in the Czech Republic using a three-parameter log-logistic distribution with the density

$$f(y;\tau,\sigma,\delta)=\begin{cases}\frac{\tau}{\sigma}\left(\frac{y-\delta}{\sigma}\right)^{\tau-1}\left(1+\left(\frac{y-\delta}{\sigma}\right)^{\tau}\right)^{-2}, & y\geq\delta>0,\ \tau>0,\ \sigma>0,\\ 0, & \text{otherwise,}\end{cases} \tag{22}$$

where $\tau>0$ is a shape parameter, $\sigma>0$ is a scale parameter and δ is a location parameter.

Vrabec and Marek (2016) also calculated the estimates of the parameters of (22) for the data of 2^{nd} quarter 2014 and obtained

$$\widehat{\tau}=4.0379,\ \widehat{\sigma}=21,687\ \text{ and }\ \widehat{\delta}=250. \tag{23}$$

The estimates (23) are based on aggregated data (frequencies by wage intervals with constant width 100 CZK) of roughly 2.1×10^6 observations, covering practically half of the overall relevant population. The corresponding estimated average monthly income is 24, 290 CZK (approximately 950 EUR).

The probability histogram of the data with bin width 500 (CZK), and density of the log-logistic distribution (22) with the unknown parameters replaced by their estimates (23), are presented in Figure 1. In addition to that, the corresponding sample distribution function is presented in Figure 2. Both the histogram and the sample distribution function were constructed from the same aggregated data from the 2^{nd} quarter 2014 used for estimation of parameters of the model. Point out that all calculation and simulations were conducted by the statistical freeware R, version 3.5.1, see R Core Team (2021).

It is interesting to look at both the lower and upper sample quantiles of the data used. While 7 000 CZK corresponds to the 0.0003 sample quantile, 8 000 CZK corresponds to the 0.01 sample quantile, whis is the reason why we set $m=7\,000$. Analogously, 40 000 CZK corresponds to the 0.91 sample quantile, 60 000 CZK to the 0.97 sample quantile and, finally, 80 000 CZK to the 0.98 sample quantile, see Figure 2.

It is obvious from Figure 1 that the original data are highly skewed. Therefore, it is not surprising that the mean relative concentration measure $\Gamma_{Y,M}=0.198$ is close to its attainable maximum, so that the estimator $t^{HT,R}_{\alpha,(m,M)}$ based on the knowledge of Υ_i's and "almost-optimal" choice of the parameter $\alpha\approx3\Gamma_{Y,M}$ should have smaller variance than $t^{HT,R}_{(m,M)}$ (corresponding to $\alpha=0$). Moreover, it follows from (4) that the variance of the estimators using the suggested RRTs will be higher than for the Horvitz-Thompson's estimator based on the nonrandomized data. All this is confirmed by our simulations, compare the results of Tables 2 – 4.

Neither the real population nor the real sample is available to us, because files with microdata from ISPV survey are not available to researchers. Therefore, the populations U are generated using the model wage distribution (log-logistic). More precisely, 1000 replications of populations sized $N=200$, or $N=400$, are simulated from model (22), in which the unknown parameters have been replaced with their estimates (23), using the package *flexsurv*, see Jackson (2016). Let us point out that the population sizes $=200$ and $N=400$ are commonly used sizes of a stratum in business statistics or surveyed community (village, group of students, etc.). It is worth to emphasize that the simulation results virtually do not change after 100 replications of the population; the differences begin at the third significant digit.

Moreover, 1000 replications from the log-logistic distribution are generated using the package *flexsurv*, see Jackson (2016). Point out that the simulation results virtually do not change after 100 replications of the population; the differences begin at the third significant digit. All simulations and calculations are conducted by statistical freeware R, version 3.5.1, see R Core Team (2021).

From each replication of the population, we draw, without replacement, 1000 random samples of the size $n = 20$, or $n = 50$. Such sample sizes are standard for separate strata in business sampling surveys, and also in the social statistical surveys, such as the EU Statistics of Income Living Condition. Let us take a closer look at average sample size per stratum in more detail. In such a survey, for a medium sized country like the Czech Republic with a population of 10^7 inhabitants and approximately $4.3 \cdot 10^6$ households, the samples approximately include 9 500 households surveyed in a two-dimensional stratification (region and size of municipality), giving $78 \times 4 = 312$ strata. The average sample size is then about 30 per stratum. In EU-SILC, detailed results are presented for eight income groups, leading on average to the population size of approximately $N = 1.25 \cdot 10^6$ inhabitants per one income group. The setting of the simulation was based on the real sample and population sizes of the EU-SILC of a medium size EU country. For a more detailed description of the stratification, strata, sample sizes, and sampling design, see GESIS (2016).

For each sample, both t_Y and $\overline{t}_Y$ are estimated using the techniques described in Section 2. Estimates of the total mean values, instead of population totals, are presented to enable a more easy comparison between the results obtained for populations with different sizes N and different sample sizes n.

In simulations, we are especially interested in the impact of "tuning parameters" m, M, T, α and α_{opt} on estimates. Taking into account the type and nature of the data that we simulate, we set the parameters as described in Table 1. The values of α_{opt} were set using the formulae for the optimal variance described in Section 2. Other parameters were chosen with regard to our experience, in particular, the monthly salary that can be perceived to be high. Since practically all available data are larger than 7 000 CZK, we set the lower bound of the interval for generating pseudorandom numbers Υ_i to $m = 7\,000$.

The results are summarized[4] in Tables 2 – 4 and in Figure 3 – 5. They show that for larger population size N and larger sample sizes n the accuracy improves substantially. The original proposal without knowledge of pseudorandom numbers seems to be also promising for real life applications. Even the method of switching questions might be applicable for large samples from large populations if prior information is available. However, more simulations using different shapes of population distributions are needed to support these hypotheses.

The reason for the lower standard deviation of $\overline{t}^{HT,R}_{\alpha,(m,M)}$, and especially $\overline{t}^{HT,R}_{\alpha_{opt},(m,M)}$, compared to $\overline{t}^{HT,R}_{(m,M)}$ and $\overline{t}^{HT,R}_{T,(m,M)}$ is that these estimators efficiently use the information on the generated numbers of Υ. Recall that we used the moment plug-in estimate for the optimal value of α.

As expected, the values of variance of the suggested estimators are higher than those of Horvitz-Thompson's estimator based on the non-randomized data. The precision of our basic proposal is practically acceptable because, according to simulations, the corresponding sample standard deviation of the estimates increased by a mere 60 % in comparison with the Horvitz-Thompson estimate for $M = 60\,000$. This result is quite reasonable, taking into account that Y is a very sensitive variable and high nonresponse (even 50 % and more

[4] In Tables 2 – 4 both the sample averages (means) and sample standard deviation (sd) of the estimates from the simulations are presented. For simplicity, we omit "HT" in the descriptions of the estimators analyzed in all figures and tables because all the estimators we compare here are of the Horvitz – Thompson's type.

in everyday practice) for direct questioning. However, note that the modification using knowledge of the values of Υ_i leads to a substantial reduction in variance. Thus, while mildly relaxing respondents' privacy on the one hand but still keeping secret the true response because the true value of the sensitive variable is never reported, this modification provides estimates whose precision is comparable with directly surveying under zero nonresponse. On the other hand, the high variability of the estimates, even the presence of negative estimates for the mean wages, shows that the modification of the switching questions described in Section 2.3 is only a theoretical exercise and cannot be recommended for practical use. Its improvement remains an open question.

Comparing in all tables the simulation results for optimal value α_{opt} of the parameter α and fixed values $\alpha = 0.75$, we see that the mean has practically not changed; however, the expected decrease occurs in the variability of the estimate. This decrease of approximately 9% of the standard deviation (sd) shows that it pays "to tune up" the procedure and its parameters according to the given problem and available data.

Both the results of Section 2.1 and the simulations show that the variance of the estimators can be greatly reduced by choice of bounds m and M. We see that for low values of the upper bound $M = 40\,000$ the proposed estimators are competitive even with the Horvitz-Thompson estimator. It follows from (15) that approximately unbiased estimators with low variance can be constructed if we use prior information on population quantiles for choice of bounds m and M. The optimal choice of bounds with respect to the minimization of the mean square error is a field of further research.

CONCLUSIONS

The paper introduces a new randomized response model and two variants of it, intended to gather information on a (positive) sensitive quantitative variable and to estimate the population total (population mean). The idea underlying the proposal is seemingly very easy and, unlike many scrambled response methods present in the literature, does not require demanding arithmetic operations from the respondents nor the use of complicated randomization devices.

It possesses three attractive properties, namely:

1. Although a quantitative estimate is the final end, the respondent is only asked for a qualitative response.
2. It is simple to use.
3. It provides a high level of anonymity to the respondent.

In the first model, respondents are first asked to generate a random number (a sort of random threshold) from a continuous uniform distribution. Then, without revealing the generated number to the interviewer, the survey participants are asked to declare whether the true value of the sensitive variable is greater than the generated number. Under this model, the privacy of the respondents is completely protected. The two variants of the model discuss the case where the generated number is also known to the interviewer, and therefore privacy is less protected. Consequently, the use of the two variants in real analyses is not recommended, since they are prone to produce misreporting and untruthful response. They have a value only from a theoretical point of view.

A disadvantage of the discussed method may, for some respondents, be a feeling of infringement on their privacy due to an extrinsic device/technique being used for generating random numbers. This problem is mainly psychological in nature and can, at least partially, be resolved by a proper explanation of the approach of the interviewer. Unfortunately, all currently used RRT procedures suffer, to some extent, from the same problem, see the thorough discussion in Chaudhuri (2017), Chaudhuri and Christofides (2013), among others.

For all suggested RRT procedures, we show their unbiasedness and derive the corresponding variance for the Horvitz-Thompson's type estimator under simple random sampling without replacement. The optimal values of the tuning parameters that enable us to minimize the variance of the suggested procedures are also discussed.

As a technical tool, two auxiliary measures are proposed. With the aid of them we can explain why and especially how the suggested RRTs increase the variance of the estimators of t_Y and $\overline{t}_Y$ for symmetrical distributions, distributions closely concentrated around their centers, or uniform distribution.

ACKNOWLEDGEMENTS

The work of the first author was partially supported by GA ČR under the Grant number P403/22/19353S. The work of the third author was prepared under Institutional Support to Long-Term Conceptual Development of Research Organization, the Faculty of Informatics and Statistics of the University of Economics, Prague. The authors are grateful to the associated editor and two unknown reviewers for their valuable comments that considerably improved the contents of this paper.

REFERENCES

ANTOCH, J., JANSSEN, P. (1989). Nonparametric regression M-quantiles. *Statistical and Probability Letters*, 8: 355–362. <https://doi.org/10.1016/0167-7152(89)90044-8>.

ARNAB, R. (1995). Optimal estimation of a finite population total under randomized response surveys. *Statistics*, 27: 175–180. <https://doi.org/10.1080/02331889508802520>.

ARNAB, R. (1998). Randomized response surveys. Optimum estimation of a finite population total. *Statistical Papers*, 39: 405–408. <https://doi.org/10.1007/BF02927102>.

ARNAB, R. (2017). *Survey Sampling, Theory and Applications.* London: Elsevier. ISBN 978-0-81148-1.

BOURKE, P., DALENIUS, T. (1976). Some new ideas in the realm of randomized inquiries. *Int. Statistical Review*, 44: 219–221. <https://doi.org/10.2307/1403280>.

BRICK, M. (2013). Unit nonresponse and weighting adjustments: A critical review. *J. Official Statistics*, 29: 329–353. <https://doi.org/10.2478/jos-2013-0026>.

CHAUDHURI, A. (1987). Randomized response surveys of a finite population: A unified approach with quantitative data. *J. Statistical Planning and Inference*, 15: 157–165. <https://doi.org/10.1016/0378-3758(86)90094-7>.

CHAUDHURI, A. (2017). *Randomized Response and Indirect Questioning Techniques in Surveys.* New York: Chapman and Hall/CRC. ISBN 978-11-3811542-2.

CHAUDHURI, A., CHRISTOFIDES, T. (2013). *Indirect Questioning in Sample Surveys.* Heidelberg: Springer. ISBN 978-3642-36275-0.

CHAUDHURI, A., MUKERJEE, R. (1988). *Randomized Response, Theory and Techniques.* New York: Marcel Dekker.

CHAUDHURI, A., CHRISTOFIDES, T., RAO, C. (2016). *Handbook of Statistics 34. Data Gathering, Analysis and Protection of Privacy Through Randomized Response Techniques.* Amsterdam: Elsevier. ISBN 978-0444-63570-9.

DALENIUS, T., VITALE, R. (1979). A new randomized response design for estimating the mean of a distribution. In: JUREČKOVÁ, J., (eds.) *Contributions to Statistics*, 54–59. Praha: Academia.

ERIKSSON, S. (1973). A new model for randomized response. *Int. Statistical Review*, 41: 101–113. <https://doi.org/10.2307/1402791>.

FOX, J. (2016). *Randomized Response and Related Methods: Surveying Sensitive Data*, 2nd Ed. London: Sage. ISBN 978-1483-38103-9. <https://doi.org/10.4135/9781506300122>.

GESIS (2016). *EU-SILC 2016, Metadata for Official Statistics.* Mannheim: Gesis Missy. <https://www.gesis.org/en/missy/metadata/EU-SILC/2016/>.

GJESTVANGA, C., SINGH, R. (2009), An improved randomized response model: Estimation of mean. *J. Applied Statistics*, 36: 1361–1367. <https://doi.org/10.1080/02664760802684151>.

GREENBERG, B. (1971). Application of the randomized response technique in obtaining quantitative data. *J. American Statistical Association*, 66: 243–250. <https://doi.org/10.1080/01621459.1971.10482248>.

HORVITZ, D., THOMPSON, D. (1952). A generalization of sampling without replacement from a finite universe. *J. American Statistical Association*, 47: 663–685. <https://doi.org/10.1080/01621459.1952.10483446>.

ISO (2010). *ISO 28640:2010 (en). Random variate generation methods.* Geneve: ISO. <https://www.iso.org/obp/ui/#iso:std:42333:en>.

JACKSON, C. (2016). flexsurv: A platform for parametric modelling in R. *J. of Statistical Software*, 70: 1–33. <https://cran.r-project.org/web/packages/flexsurv/index.html>.

KIRCHNER, A. (2015). Validating sensitive questions: A comparison of survey and register data. *J. Official Statistics*, 31: 31–59. <https://doi.org/10.1515/jos-2015-0002>.

PFEFFERMANN, D., RAO, R.C. (2009a). *Handbook of Statistics 29A. Sample Surveys: Design, Methods and Application.* Amsterdam: Elsevier. ISBN 978-0444-53124-7.

PFEFFERMANN, D., RAO, R.C. (2009b). *Handbook of Statistics 29B. Sample Surveys: Inference and Analysis.* Amsterdam: Elsevier. ISBN 978-0444-53438-5.

R CORE TEAM (2021). *R: A language and environment for statistical computing.* Austria, Vienna: R Foundation for Statistical Computing. <https://www.R-project.org>.

SÄRNDAL, C., LUNDSTRÖM, S. (2005). *Estimation in Surveys with Nonresponse.* Chichester: J. Wiley and Sons. ISBN 978-0470-01133-1.

SÄRNDAL, C., SWENSSON, B., WRETMAN, J. (1992). *Model Assisted Survey Sampling.* Heidelberg: Springer. ISBN 978-0387-40620-6.

STEEH, C. (2001). Are they really as bad as they seem? Nonresponse rates at the end of the twentieth century. *J. Official Statistics*, 17: 227–247.

STOOP, I. (2005). *The Hunt for the Last Respondent: Nonresponse in Sample Surveys.* The Hague: Social and Cultural Planning Office of the Netherlands. ISBN 90-377-0215-5.

TIAN, G.-L., TANG, M.-L. (2014). *Incomplete Categorical Data Design: Non-Randomized Response Techniques for Sensitive Questions in Surveys.* Boca Raton: Chapman & Hall/CRC. ISBN 978-1439-85533-1.

TILLÉ, Y. (2006). *Sampling Algorithms.* New York: Springer. ISBN 978-0387-30814-2.

TILLÉ, Y. (2020). *Sampling and Estimation from Finite Populations.* New York: J. Wiley and Sons. ISBN 978-0470-68205-0.

TRAPPMANN, M. (2014). A new technique for asking quantitative sensitive questions. *J. Survey Statistics and Methodology*, 2: 58–77. <https://doi.org/10.1093/jssam/smt019>.

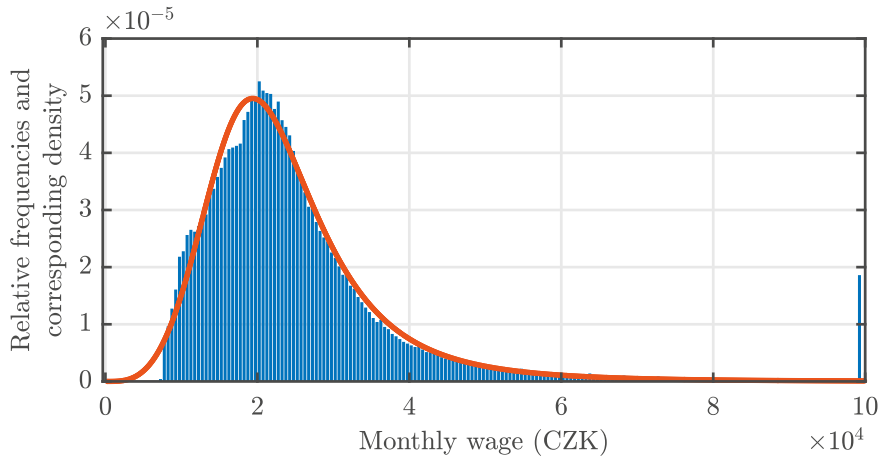
VRABEC, M., MAREK, L. (2016). Model of distribution of wages. *AMSE 2016, 19th Symp. Applications of Mathematics and Statistics in Economics*, Banská Štiavnica, 378–396. <https://amsesite.wordpress.com>.

WARNER, S. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *J. American Statistical Assoc.*, 60: 63–69. <https://doi.org/10.2307/2283137>.

WU, C., THOMPSON, M. (2020). *Sampling Theory and Practice.* Basel: Springer.

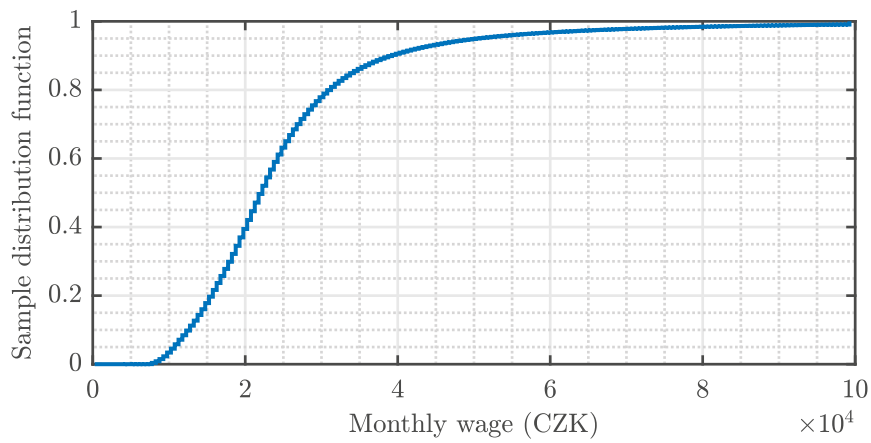
ANNEX

Figure 1 Probability histogram of monthly wages in the Czech Republic in the 2nd quarter of 2014, and the density (in red) of approximating model (22) with the parameters estimated by (23)



Source: Own construction

Figure 2 The sample distribution function of monthly wages in the Czech Republic in the 2nd quarter of 2014



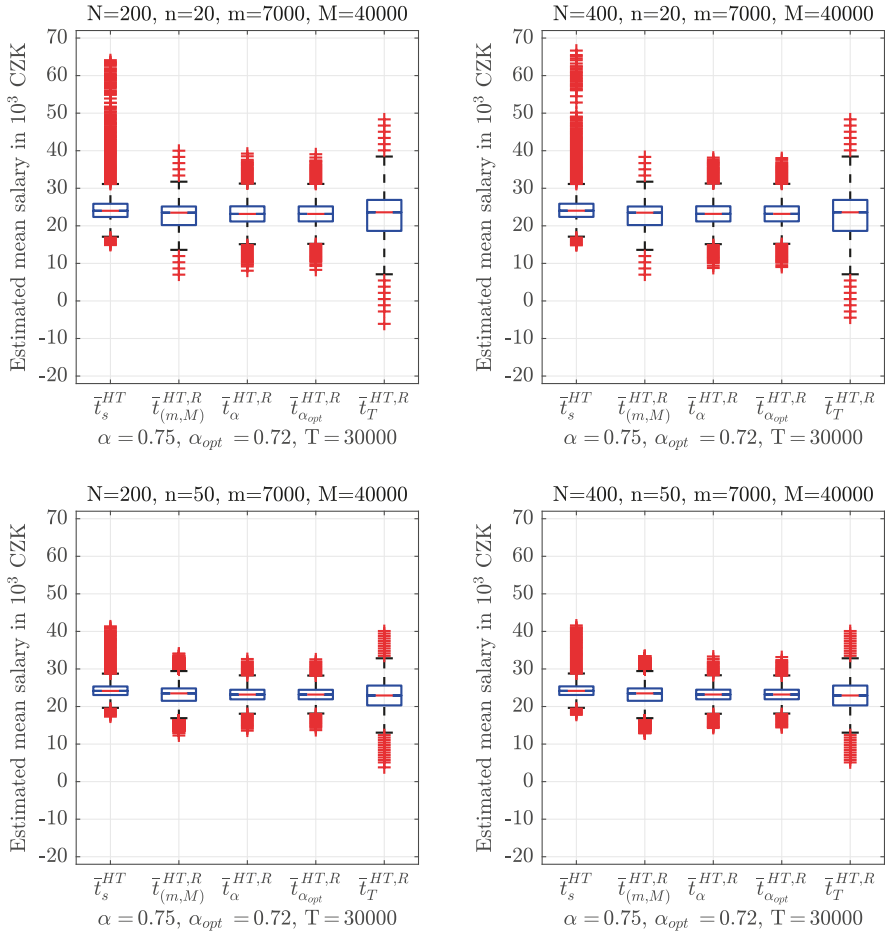
Source: Own construction

Table 1 Choice of tuning parameters for the simulations

m	M	T	α	α_{opt}
7 000	40 000	30 000	0.75	0.72
7 000	60 000	45 000	0.75	0.59
7 000	80 000	45 000	0.75	0.52

Source: Own construction

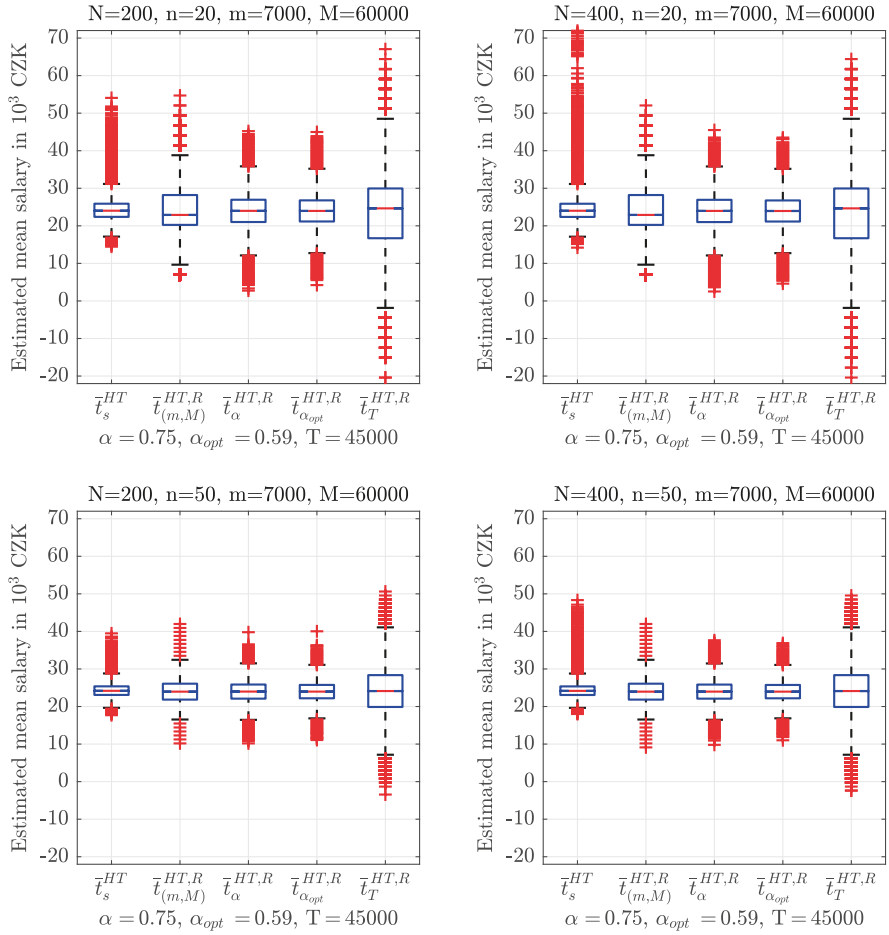
Figure 3 Behavior of considered estimators applied to different population and sample sizes



Parameters of the simulation, population N and sample sizes n : $(m, M) = (7\,000; 40\,000)$, $T = 30\,000$, $\alpha = 0.75$ and $\alpha_{opt} = 0.72$. To increase readability, we use $\overline{t}_{\alpha}^{HT,R}$, $\overline{t}_{\alpha_{opt}}^{HT,R}$ and $t_T^{HT,R}$ instead of $\overline{t}_{\alpha,(m,M)}^{HT,R}$, $\overline{t}_{\alpha_{opt},(m,M)}^{HT,R}$ and $t_{T,(m,M)}^{HT,R}$ in description of boxplots.

Source: Own construction

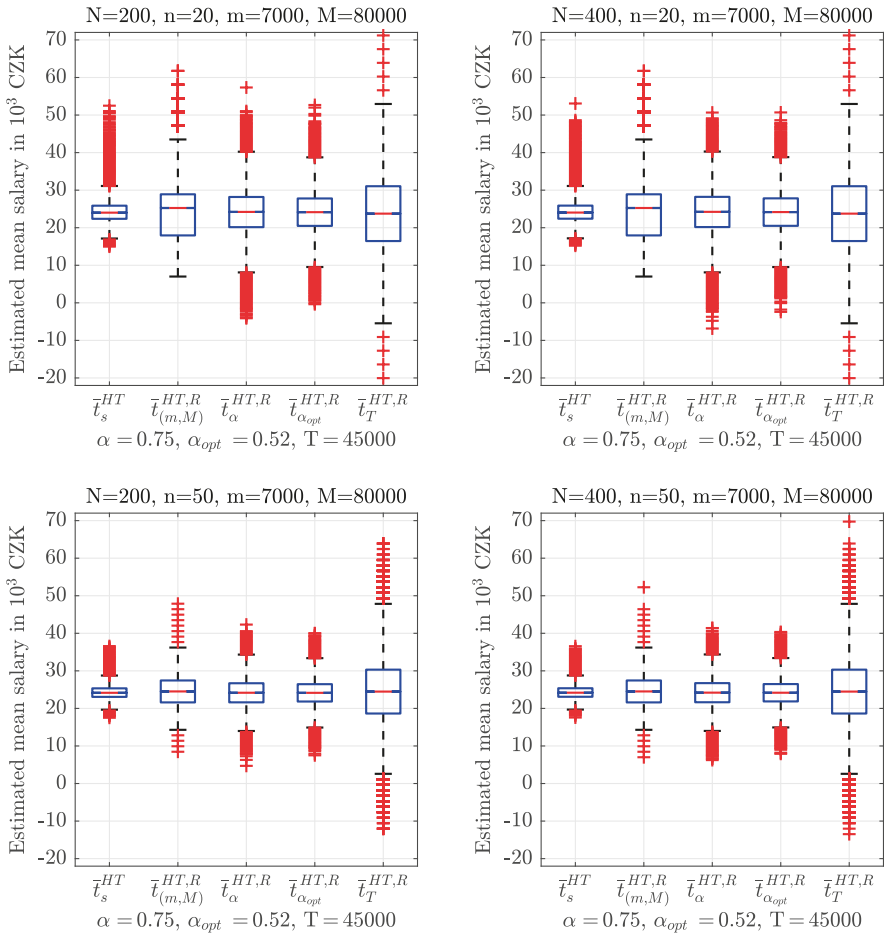
Figure 4 Behavior of considered estimators applied to different population and sample sizes



Parameters of the simulation, population N and sample sizes $n : (m, M) = (7\,000; 60\,000)$, $T = 45\,000$, $\alpha = 0.75$ and $\alpha_{opt} = 0.59$. To increase readability, we use $\overline{t}_{\alpha}^{HT,R}$, $\overline{t}_{\alpha_{opt}}^{HT,R}$ and $t_T^{HT,R}$ instead of $\overline{t}_{\alpha,(m,M)}^{HT,R}$, $\overline{t}_{\alpha_{opt},(m,M)}^{HT,R}$ and $t_{T,(m,M)}^{HT,R}$ in description of boxplots.

Source: Own construction

Figure 5 Behavior of considered estimators applied to different population and sample sizes



Parameters of the simulation, population N and sample sizes n : $(m, M) = (7\,000; 80\,000)$, $T = 45\,000$, $\alpha = 0.75$ and $\alpha_{opt} = 0.52$. To increase readability, we use $\overline{t}_{\alpha}^{HT,R}$, $\overline{t}_{\alpha_{opt}}^{HT,R}$ and $t_T^{HT,R}$ instead of $\overline{t}_{\alpha,(m,M)}^{HT,R}$, $\overline{t}_{\alpha_{opt},(m,M)}^{HT,R}$ and $t_{T,(m,M)}^{HT,R}$ in description of boxplots.

Source: Own construction

Table 2 Numerical results of simulations

		$N = 200$		$N = 400$	
Estimator		$n = 20$	$n = 50$	$n = 20$	$n = 50$
$\overline{t}_s^{HT}$	mean	24.270	24.272	24.287	24.288
	sd	2.782	1.757	2.773	1.758
$\overline{t}_{(m,M)}^{HT,R}$	mean	23.189	23.192	23.203	23.205
	sd	3.687	2.333	3.690	2.336
$\overline{t}_{\alpha,(m,M)}^{HT,R}$	mean	23.192	23.194	23.206	23.207
	sd	3.000	1.897	3.001	1.902
$\overline{t}_{\alpha_{opt},(m,M)}^{HT,R}$	mean	23.192	23.194	23.206	23.207
	sd	2.965	1.875	2.966	1.880
$\overline{t}_{T,(m,M)}^{HT,R}$	mean	23.185	23.189	23.199	23.202
	sd	6.066	3.836	6.068	3.837

The mean estimated salaries (in 10^3 CZK) and the corresponding sample standard deviations (in 10^3 CZK) for different population sizes N and sample sizes n. Random numbers Υ_i are generated from the uniform distribution on the interval $[m, M] = [7\,000; 40\,000]$, $T = 30\,000$, $\alpha = 0.75$, $\alpha_{opt} = 0.72$, 1000 simulated populations, 1000 replications of each. Means and standard deviations (sd) were averaged over 1000×1000 random samples.

Table 3 Numerical results of simulations

		$N = 200$		$N = 400$	
Estimator		$n = 20$	$n = 50$	$n = 20$	$n = 50$
$\overline{t}_s^{HT}$	mean	24.297	24.301	24.288	24.290
	sd	2.773	1.758	2.813	1.779
$\overline{t}_{(m,M)}^{HT,R}$	mean	23.983	23.984	23.965	23.974
	sd	5.530	3.501	5.529	3.495
$\overline{t}_{\alpha,(m,M)}^{HT,R}$	mean	23.974	23.976	23.956	23.965
	sd	4.401	2.786	4.398	2.780
$\overline{t}_{\alpha_{opt},(m,M)}^{HT,R}$	mean	23.976	23.977	23.958	23.967
	sd	4.164	2.637	4.161	2.631
$\overline{t}_{T,(m,M)}^{HT,R}$	mean	23.991	23.992	23.973	23.982
	sd	9.066	5.729	9.067	5.726

The mean estimated salaries (in 10^3 CZK) and the corresponding standard deviations (in 10^3 CZK) for different population sizes N and sample sizes n. Random numbers Υ_i are generated from the uniform distribution on the interval $[m, M] = [7\,000; 60\,000]$, $T = 45\,000$, $\alpha = 0.75$, $\alpha_{opt} = 0.59$, 1000 simulated populations, 1000 replications of each. Means and standard deviations (sd) were averaged over 1000×1000 random samples.

Table 4 Numerical results of simulations

		$N = 200$		$N = 400$	
Estimator		$n = 20$	$n = 50$	$n = 20$	$n = 50$
$\overline{t}_s^{HT}$	mean	24.275	24.273	24.299	24.299
	sd	2.765	1.739	2.753	1.737
$\overline{t}_{(m,M)}^{HT,R}$	mean	24.138	24.140	24.158	24.168
	sd	6.911	4.372	6.921	4.378
$\overline{t}_{\alpha,(m,M)}^{HT,R}$	mean	24.145	24.146	24.165	24.174
	sd	5.962	3.770	5.950	3.767
$\overline{t}_{\alpha_{opt},(m,M)}^{HT,R}$	mean	24.143	24.145	24.163	24.173
	sd	5.404	3.417	5.398	3.417
$\overline{t}_{T,(m,M)}^{HT,R}$	mean	24.136	24.137	24.156	24.165
	sd	13.018	8.236	13.036	8.244

Numerical results of simulations. The mean estimated salaries (in 10^3 CZK) and the corresponding standard deviations (in 10^3 CZK) for different population sizes N and sample sizes n. Random numbers Υ_i are generated from the uniform distribution on the interval $[m, M] = [7\,000; 80\,000], T = 45\,000$, $\alpha = 0.75$, $\alpha_{opt} = 0.52$, 1000 simulated populations, 1000 replications of each. Means and standard deviations (sd) were averaged over 1000×1000 random samples.

Source of Tables 2 – 4: Own construction