

Multiple sensitive estimation and optimal sample size allocation in the item sum technique

Pier Francesco Perri¹  | María del Mar Rueda García² | Beatriz Cobo Rodríguez²

¹Department of Economics, Statistics and Finance, University of Calabria. Via P. Bucci, 87036 Arcavacata di Rende, Italy

²Department of Statistics and Operational Research, University of Granada. Campus Universitario Fuentenueva, 18071 Granada, Spain

Correspondence

Pier Francesco Perri, Department of Economics, Statistics and Finance, University of Calabria. Via P. Bucci, 87036 Arcavacata di Rende, Italy.

Email: pierfrancesco.perri@unical.it

Abstract

For surveys of sensitive issues in life sciences, statistical procedures can be used to reduce nonresponse and social desirability response bias. Both of these phenomena provoke nonsampling errors that are difficult to deal with and can seriously flaw the validity of the analyses. The item sum technique (IST) is a very recent indirect questioning method derived from the item count technique that seeks to procure more reliable responses on quantitative items than direct questioning while preserving respondents' anonymity. This article addresses two important questions concerning the IST: (i) its implementation when two or more sensitive variables are investigated and efficient estimates of their unknown population means are required; (ii) the determination of the optimal sample size to achieve minimum variance estimates. These aspects are of great relevance for survey practitioners engaged in sensitive research and, to the best of our knowledge, were not studied so far. In this article, theoretical results for multiple estimation and optimal allocation are obtained under a generic sampling design and then particularized to simple random sampling and stratified sampling designs. Theoretical considerations are integrated with a number of simulation studies based on data from two real surveys and conducted to ascertain the efficiency gain derived from optimal allocation in different situations. One of the surveys concerns cannabis consumption among university students. Our findings highlight some methodological advances that can be obtained in life sciences IST surveys when optimal allocation is achieved.

KEYWORDS

complex sampling, Horvitz–Thompson estimator, indirect questioning methods, sensitive research

1 | INTRODUCTION

Studies in life and social sciences addressing highly personal, embarrassing, stigmatizing, threatening, or even incriminating issues often yield unreliable estimates of unknown characteristics of the population under study, due to nonresponse (unit-nonresponse or item-nonresponse) and socially desirable responding. In particular, social desirability bias, that is the desire to make a favorable impression on others, poses a significant threat to the validity of self-reports in “sensitive research” as well described in Dickson-Swift, James, and Liamputtong (2008).

Refusal to answer and false answers constitute nonsampling errors that are difficult to deal with and can seriously flaw the quality of the collected data, thus jeopardizing the usefulness of subsequent analyses including statistical inference of unknown characteristics of the population under study. Although these errors cannot be totally avoided, they may be mitigated by enhancing respondents' cooperation. Since the decision to cooperate fully and honestly greatly depends on how survey participants perceive their privacy being disclosed, survey modes that ensure respondents' anonymity or, at least, a high degree of confidentiality, may

go some way to improving cooperation and, consequently, ensure more reliable information on sensitive topics than that derived from direct questioning.

In recent years, indirect questioning survey modes have gained popularity in many research fields, mostly falling in the life and social sciences, as effective methods for eliciting truthful responses to sensitive questions while guaranteeing respondents' privacy. In general, this nonstandard survey approach encourages greater cooperation from respondents and reduces the motivation to falsely report their attitudes. The approach obeys the principle that no direct question is posed to survey participants and, then, there is no need for respondents to openly reveal if they are actually engaged in sensitive behaviors. In this way, privacy is protected since answers remain confidential to the respondents and, consequently, their true status remains uncertain and undisclosed to both the interviewer and the researcher. Nonetheless, although the individual information provided by the respondents cannot be used to know their true sensitive status, the information gathered for all the survey participants can be profitably used to make inference on certain parameters of interest of the population under study, usually the prevalence of a sensitive behavior, its frequency or the mean/total of a sensitive quantitative variable.

The indirect questioning strategies may be classified in three different groups: the randomized response technique, the item count technique (ICT), and the nonrandomized response technique. All the approaches have produced a considerable literature and attracted the interest of health, cognitive and behavioral psychologists, epidemiologists, health-care operators, researchers engaged in organizing, managing and conducting sensitive studies, as well as policy-makers committed in formulating effective diseases and mental disorders control measures and promoting public intervention programs to gauge progress toward improving the behavioral health of a state.

For a comprehensive review of the topic, interested readers are referred to Fox and Tracy (1986), Chaudhuri and Mukerjee (1988), Chaudhuri (2011), Chaudhuri and Christofides (2013), Tian and Tang (2014). Useful and detailed studies on recent methodological advances, more complex estimation problems and new challenges may be found, among others, in Arcos, Rueda, and Singh (2015), Barabesi, Diana, and Perri (2013, 2015), Diana and Perri (2011), Fox, Entink, and Avetisyan (2014), Glynn (2013), Groenitz (2014), Hoffmann and Musch (2016), Hoffmann, Diedenhofen, Verschuere, and Musch (2016), Husain, Shabbir, and Shabbir (2015), Ibrahim (2016), Imai (2011), Imai, Park, and Greene (2015), Liu and Tian (2013), Moshagen, Hilbig, Erdfelder, and Moritz (2014), Nepusz, Petróczi, Naughton, Epton, and Norman (2014), Perri and van der Heijden (2012), Petróczi et al. (2011), Rueda, Cobo, and Arcos (2016), Tsuchiya (2005), Ulrich, Schörter, Striegel, and Simon (2012), Wu and Tang (2016).

Various indirect questioning techniques have been experienced in different branches of life sciences. In particular, these methods have been mainly applied to estimate prevalence of discriminating or embarrassing behaviors in epidemiological and medical studies. Some recent contributions, although not exhaustive, cover a great variety of topics. For instance: the measure of the impact of HIV/AIDS infection in Botswana (Arnab & Singh, 2010); the assessment of sensitive health-risk behaviors in HIV/AIDS positive individuals (Arentoft et al., 2016); the assessment of permissive sexual attitudes and high-risk sexual behaviors to reduce the transmission and acquisition of sexually transmitted infections and HIV/AIDS (De Jong, Pieters, & Stremersch, 2012; Starosta & Earleywine, 2014; Geng, Gao, Ruan, Yu, & Zhou, 2016; Kazemzadeh, Shokoohi, Baneshi, & Haghdoost, 2016); patterns of condom use among university students for HIV/AIDS control programs (Safiri, 2016; Vakilian, Mousavi, Keramat, & Chaman, 2016); the prevalence of sexual behaviors such as extradyadic sex (Tu & Hsieh, 2017), commercial sex among homosexual men (Chen et al., 2014) and sexual assault (Krebs et al., 2011); the use of drug, and athletic, cognitive, and mood performance-enhancing substances (Striegel, Ulrich, & Simon, 2010; Petróczi et al., 2011; Dietz et al., 2013; Franke et al., 2013; James, Nepusz, Naughton, & Petróczi, 2013; Nakhaee, Pakravan, & Nakhaee, 2013; Stubbe, Chorus, Frank, de Hon, & van der Heijden, 2013; Shamsipour et al., 2014; Khosravi et al., 2015; Cobo, Rueda, & López-Torrecillas, 2016); smoking behavior validation studies (Fox, Avetisyan, & van der Palen, 2013); dental hygiene habits of Chinese college students (Moshagen, Musch, Ostapczuk, & Zhao, 2010); farmers' transgressionary behaviors and prevalence of animal diseases such as sheep scab in Wales (Cross, Edwards-Jones, Omed, & Williams 2010), African swine fever in Madagascar (Randrianantoandro, Kono, & Kubota, 2015), or foot and mouth disease-infected animals in Sri Lanka (Gunarathne, Kubota, Kumarawadu, Karunagoda, & Kon, 2016); estimation of the prevalence of induced abortion (Oliveras & Letamo, 2010; Moseson et al., 2015; Perri, Pelle, & Stranges, 2016); ecological and biological conservation issues including estimation of illegal bushmeat hunting (Nuno, Bunnefeld, Naiman, & Milner-Gulland, 2013; Conteh, Gavin, & Solomon, 2015), illegal fishing (Blank & Gavin, 2009; Arias & Sutton, 2013), and unauthorized natural resources use (Harrison, Baker, Twinamatsiko, & Milner-Gulland, 2015).

This article focuses on a recent variant of the ICT conceived to deal with quantitative sensitive variables. We propose some methodological advances that can be useful in life sciences when multiple sensitive issues are to be investigated, and reliable and accurate estimates of usually underreported characteristics are to be produced.

The ICT has recently attracted much attention among applied researchers. This method, also known as the list experiment or the unmatched count technique, was originally proposed by Miller (1984) for binary variables to estimate the prevalence of a

stigmatizing behavior within the population. Without loss of generality, respondents are asked directly about their own sensitive behavior and, at the same time, about a number of innocuous behaviors. In the standard setting, the method requires the selection of two samples: a reference sample that receives a short list (SL) of items on questions only about innocuous behaviors, and a treatment sample that receives a long list (LL) containing the innocuous items in the SL-sample and a sensitive question. Units selected in the two samples are asked to report the total number of items that apply to them without revealing which item applies individually.

The ICT is used in surveys that require the study of a qualitative variable. Nonetheless, many practical situations may deal with sensitive variables that are quantitative in nature. To address this situation, Chaudhuri and Christofides (2013) proposed a generalization of the ICT that can be used to estimate the mean (or the total) of a quantitative variable. Trappmann, Krumpal, Kirchner, and Jann (2014) called this variant the item sum technique (IST) and used it in a survey to estimate the amount of undeclared work in Germany. The IST works in a similar way to the ICT and offers a promising tool for dealing with sensitive issues. Nonetheless, some methodological challenges, conceptually inherited from the ICT, remain to be overcome in order to successfully use the technique in applied research. The purpose of the present article is to address these challenges. In particular, two open and unresolved issues are discussed. The first pertains the reduction of the statistical burden when multiple sensitive items are to be investigated and estimates of certain characteristics are required. This situation occurs frequently in real studies where researchers must incorporate $Q \geq 2$ sensitive questions in their surveys. Three different approaches are considered in the article, and pros and cons highlighted. The first two techniques require that sampled units participate in Q distinct IST surveys, one for each sensitive item. The first method is time-consuming and costly since requires the selection of $2Q$ samples, the second instead requires Q samples but burdens the surveyed participants. A third viable alternative, which requires the selection of $Q + 1$ samples and acts as a trade-off between the first two approaches, is therefore proposed and its performance investigated on a number of simulation experiments based on real data.

The second, but not less important, problem we consider is how to split the total sample size into the LL-sample and the SL-sample. A simple solution would be to allocate the same number of units to each sample, irrespective of the variability of the items in the two lists. Although intuitive and easy to implement, this basic solution is inefficient because estimates may be affected by high variability. A possible alternative, discussed in the article, would be to achieve optimal sample size allocation by minimizing the variance of the IST estimates under a budget constraint. This possibility is first formalized and discussed under a generic sample design and, then, results are particularized to the simple random sampling and the stratified sampling designs. Optimal allocation results are finally extended to the multiple sensitive estimation setting.

Methodological developments are integrated with an extensive simulation study aimed at investigating the performance of the proposed techniques and the related estimators under two different sampling designs and for different sample sizes. Most of the simulation study is based on the results of a real sensitive research conducted among university students in Granada (Spain) to investigate the consumption of cannabis for recreational purposes.

The rest of this paper is organized as follows: Section 2 introduces the IST under a very general sampling design. Section 3 discusses some estimation methods for multiple sensitive questions under different approaches. The problem of the optimal sample size allocation is then formulated in Section 4. Allocation is first derived for a general setting and then applied to simple random sampling without replacement and stratified sampling designs. In Section 5, a number of simulation experiments are generated from two real surveys to investigate the performance of the optimal allocation for single and multiple sensitive estimation under different scenarios. One of the surveys concerns the number of cannabis cigarettes smoked in last year by university students. Section 6 concludes the article with some final considerations.

2 | THE ITEM SUM TECHNIQUE

Consider a finite population $U = \{1, \dots, N\}$ consisting of N different and identifiable units. Let y_i be the value of the sensitive character under study, say \mathcal{Y} , for the i -th population unit. Let us suppose that the population mean $\bar{Y} = N^{-1} \sum_{i \in U} y_i$ is unknown and has to be estimated in an IST setting. In so doing, two independent samples, say s_{ll} and s_{sl} , are selected from U according to the generic sampling designs $p_{ll}(\cdot)$ and $p_{sl}(\cdot)$ with positive first- and second-order inclusion probabilities $\pi_{i(ll)} = \sum_{s_{ll} \ni i} p_{ll}(s_{ll})$, $\pi_{ij(ll)} = \sum_{s_{ll} \ni i, j} p_{ll}(s_{ll})$, $\pi_{i(sl)} = \sum_{s_{sl} \ni i} p_{sl}(s_{sl})$, and $\pi_{ij(sl)} = \sum_{s_{sl} \ni i, j} p_{sl}(s_{sl})$ with $i, j \in U$. Let $d_{i(ll)} = \pi_{i(ll)}^{-1}$ and $d_{i(sl)} = \pi_{i(sl)}^{-1}$ denote the known sampling design-basic weight for unit $i \in U$ in each sampling design.

Chaudhuri and Christofides (2013) introduced the IST in the following way: one of the samples, say s_{ll} , is confronted with a LL of items containing $G + 1$ questions of which G refer to nonsensitive characteristics and one is related to the sensitive characteristic under study. The other sample, s_{sl} , receives a SL of items that only contains the G innocuous questions present

in the LL-sample. All sensitive and nonsensitive items are quantitative in nature. Respondents in each sample are requested to report the total score of all the items applicable to them, without revealing the individual scores for the items.

Without loss of generality, let \mathcal{T} be the variable denoting the total score applicable to the G nonsensitive questions, and $\mathcal{Z} = \mathcal{Y} + \mathcal{T}$ the total score applicable to the nonsensitive questions and the sensitive question. When $G = 1$, \mathcal{T} denotes the innocuous variable and t_i its value on unit $i \in U$. Hence, the answer given by the i -th respondent will be $z_i = y_i + t_i$ if $i \in s_{ll}$ or t_i if $i \in s_{sl}$.

Under the sampling designs $p_{ll}(\cdot)$, $p_{sl}(\cdot)$, let:

$$\hat{Z} = \frac{1}{N} \sum_{i \in s_{ll}} d_{i(ll)} z_i, \quad \hat{T} = \frac{1}{N} \sum_{i \in s_{sl}} d_{i(sl)} t_i$$

be the unbiased Horvitz–Thompson (hereafter HT) estimators of $\bar{Z} = N^{-1} \sum_{i \in U} (y_i + t_i)$ and $\bar{T} = N^{-1} \sum_{i \in U} t_i$, respectively. Hence, a HT-type estimator of \bar{Y} under the IST can be readily obtained as:

$$\hat{Y} = \hat{Z} - \hat{T}. \quad (1)$$

From the unbiasedness of \hat{Z} and \hat{T} , it readily follows that the estimator \hat{Y} is unbiased for \bar{Y} . Furthermore, as long as the two samples are independent, the variance of \hat{Y} can be expressed as:

$$\begin{aligned} \mathbb{V}(\hat{Y}) &= \mathbb{V}(\hat{Z}) + \mathbb{V}(\hat{T}) = \\ &= \frac{1}{N^2} \left(\sum_{i,j \in U} \Delta_{ij(ll)} d_{i(ll)} d_{j(ll)} z_i z_j + \sum_{i,j \in U} \Delta_{ij(sl)} d_{i(sl)} d_{j(sl)} t_i t_j \right), \end{aligned} \quad (2)$$

where $\Delta_{ij(a)} = \pi_{ij(a)} - \pi_{i(a)}\pi_{j(a)}$ with $a = ll, sl$. An unbiased estimator of $\mathbb{V}(\hat{Y})$ is given by:

$$\hat{\mathbb{V}}(\hat{Y}) = \frac{1}{N^2} \left(\sum_{i,j \in s_{ll}} \check{\Delta}_{ij(ll)} d_{i(ll)} d_{j(ll)} z_i z_j + \sum_{i,j \in s_{sl}} \check{\Delta}_{ij(sl)} d_{i(sl)} d_{j(sl)} t_i t_j \right)$$

where $\check{\Delta}_{ij(a)} = \Delta_{ij(a)} / \pi_{ij(a)}$.

3 | MULTIPLE SENSITIVE ESTIMATION UNDER IST

Traditionally, indirect questioning techniques deal with one sensitive variable. However, in real surveys, the researcher may be interested in investigating more than one sensitive variable. Typical areas of inquiry include: (i) the amount of self-employment income and income from financial assets; (ii) the frequency and amount of tax evasion; (iii) the frequency, quantity, and cost of cannabis use. In general, in situations like these concerning multiple estimation of the means of $Q > 1$ quantitative sensitive variables, the implementation of the IST may be not unique and cumbersome, for various reasons. To obtain a reliable estimation, a number of solutions might be adopted. One consists in performing Q separate IST surveys, one for each sensitive item. This approach (hereafter, separate approach) requires for each item the selection of one LL-sample and one SL-sample, for a total of $2Q$ samples. In practice, however, this solution does not appear to be feasible, because it is both time-consuming and costly, and also because possible associations between variables would be lost since each IST survey is independently executed on different subjects. To overcome these problems, a single IST survey could be performed. In this case, just one LL-sample and one SL-sample are selected and respondents are asked to participate in Q separate IST experiments, one for each sensitive item. As can be readily imagined, this procedure (hereafter, all-in-one approach) imposes a heavy statistical burden on the respondents, since they must provide the required information on the single sensitive items by separately implementing the IST Q times. More specifically, each respondent belonging to the SL-sample has to answer on Q different short lists and each respondent in the LL-sample has to answer on Q different long lists. If there are many items to be investigated, the accuracy of the responses may deteriorate during the runs. Respondents may be more willing to participate and concentrate more effectively at the beginning of the process, but lose attention during the course of the survey, and possibly break the rules or drop out. If the all-in-one approach is adopted, the order of the items to be investigated, the question of reducing the statistical burden and the problem of respondent drop out must all be carefully considered in the survey design. In view of the manifest weaknesses of the separate and all-in-one approaches, we now consider a possible solution, one providing a trade-off of costs and benefits. Without loss of

generality, let us focus, initially, on two quantitative sensitive variables, \mathcal{Y}_1 and \mathcal{Y}_2 , and on one innocuous variable \mathcal{T} . We want to estimate the mean of the two variables, say \bar{Y}_1 and \bar{Y}_2 . Under this approach (hereafter, mixed approach), three independent samples are selected. For ease of notation, let us suppose that the same sampling design $p(\cdot)$ is used. Hence, let:

- (i) s_0 be a sample of size n_0 . The respondents are given a SL containing only the innocuous variable. The i_0 -th respondent provides the score t_{i_0} with $i_0 = 1, \dots, n_0$;
- (ii) s_1 be a sample of size n_1 . The respondents are given a list containing one sensitive variable, for instance \mathcal{Y}_1 , and the innocuous one. The i_1 -th respondent provides the total score $y_{1i_1} + t_{i_1}$ with $i_1 = 1, \dots, n_1$;
- (iii) s_2 be a sample of size n_2 . The respondents are given a list containing the two sensitive variables and the innocuous one. The i_2 -th respondent provides the total score $y_{1i_2} + y_{2i_2} + t_{i_2}$ with $i_2 = 1, \dots, n_2$.

Let

$$\hat{Z}_0 = \frac{1}{N} \sum_{i_0 \in s_0} \frac{t_{i_0}}{\pi_{i_0}}, \quad \hat{Z}_1 = \frac{1}{N} \sum_{i_1 \in s_1} \frac{y_{1i_1} + t_{i_1}}{\pi_{i_1}}, \quad \hat{Z}_2 = \frac{1}{N} \sum_{i_2 \in s_2} \frac{y_{1i_2} + y_{2i_2} + t_{i_2}}{\pi_{i_2}}.$$

Hence

$$\hat{Y}_1^* = \hat{Z}_1 - \hat{Z}_0$$

is the HT-unbiased estimator of \bar{Y}_1 with

$$\mathbb{V}(\hat{Y}_1^*) = \mathbb{V}(\hat{Z}_1) + \mathbb{V}(\hat{Z}_0).$$

Similarly,

$$\hat{Y}_2^* = \hat{Z}_2 - \hat{Z}_1$$

is the HT-unbiased estimator of \bar{Y}_2 with

$$\mathbb{V}(\hat{Y}_2^*) = \mathbb{V}(\hat{Z}_2) + \mathbb{V}(\hat{Z}_1).$$

This framework can be readily extended to the case of $Q \geq 2$ sensitive variables, $\mathcal{Y}_1, \dots, \mathcal{Y}_Q$, by selecting $Q + 1$ samples. With the same notation as in the case $Q = 2$, let:

$$\hat{Z}_k = \frac{1}{N} \sum_{i_k \in s_k} \frac{z_{i_k}}{\pi_{i_k}} = \frac{1}{N} \sum_{i_k \in s_k} \frac{\sum_{j=1}^Q y_{ji_k} + t_{i_k}}{\pi_{i_k}},$$

with $k = 1, \dots, Q$. Hence, the estimator

$$\hat{Y}_k^* = \hat{Z}_k - \hat{Z}_{k-1}$$

is the HT-unbiased estimator of \bar{Y}_k , $k = 1, \dots, Q$. The variance of this estimator is given by:

$$\begin{aligned} \mathbb{V}(\hat{Y}_k^*) &= \mathbb{V}(\hat{Z}_k) + \mathbb{V}(\hat{Z}_{k-1}) = \\ &= \frac{1}{N^2} \left(\sum_{i,j \in U} \Delta_{ij(k)} d_{i(k)} d_{j(k)} z_i z_j + \sum_{i,j \in U} \Delta_{ij(k-1)} d_{i(k-1)} d_{j(k-1)} z_i z_j \right), \end{aligned}$$

where, slightly changing the notation, $d_{b(a)} = \pi_{b(a)}^{-1}$ and $\Delta_{ij(a)} = \pi_{ij(a)} - \pi_{i(a)}\pi_{j(a)}$, with $b = i, j$ and $a = 1, \dots, k$. Accordingly, an unbiased estimator for $\mathbb{V}(\hat{Y}_k^*)$ follows as:

$$\hat{\mathbb{V}}(\hat{Y}_k^*) = \frac{1}{N^2} \left(\sum_{i_k, j_k \in s_k} \check{\Delta}_{ij(k)} d_{i(k)} d_{j(k)} z_{i_k} z_{j_k} + \sum_{i_{k-1}, j_{k-1} \in s_{k-1}} \check{\Delta}_{ij(k-1)} d_{i(k-1)} d_{j(k-1)} z_{i_{k-1}} z_{j_{k-1}} \right).$$

Similarly, $G > 1$ innocuous variables, say $\mathcal{T}_1, \dots, \mathcal{T}_G$, can be considered. In this case, \mathcal{T} denotes the total score of the values of the G innocuous variables and $t_{i_k} = \sum_{g=1}^G t_{gi_k}$ is the total score of the G innocuous variables for the i_k -th respondent in the k -th sample s_k .

4 | TOTAL SAMPLE SIZE ALLOCATION IN THE IST ESTIMATION

A key design decision in an IST survey is how to split the total sample into the LL-sample and SL-sample. A simple solution is to allocate the same number of units to each sample irrespective of the variability of the items in the two lists. Clearly, this intuitive and basic solution is not efficient because responses in the LL-sample are tendentially affected by high variability due to the presence of innocuous items: the larger the number of items, the higher the variability of the response and, hence, of the estimates. To the best of our knowledge, the problem of optimal allocation in the IST framework has not been considered so far. Therefore, we propose a possible solution to this problem. First, we consider the standard IST with just one sensitive variable, and assume that the total sample size n is fixed beforehand. Hence, the problem of optimal sample allocation is formulated as one of determining the LL-sample and SL-sample sizes, n_{ll} and n_{sl} , in such a way as to minimize the variance of \hat{Y} subject to a fixed cost C .

4.1 | Allocation under a generic sampling design

Suppose that an IST design has been decided upon. Let n be the sample size of the IST design, or the expected sample size if the sampling design is not of a fixed size. To estimate the population mean \bar{Y} , the HT-estimator defined in (1) is considered. Before selecting the sample, the sample sizes n_{ll} and n_{sl} must be determined. We provide a solution to this allocation problem for the case in which the sampling designs $p_{ll}(\cdot)$ and $p_{sl}(\cdot)$ provide a variance of the estimator that can be formulated as:

$$\mathbb{V}(\hat{Y}) = \frac{A_z}{n_{ll}} + \frac{A_t}{n_{sl}} + B, \quad (3)$$

where the terms A_z , A_t , and B do not depend on n_{ll} and n_{sl} . The simple random sampling and the stratified random sampling designs meet this requirement.

Let c_0 represent the fixed overhead cost of the survey, and $c_{ll} > 0$ and $c_{sl} > 0$ be the costs of surveying one element in s_{ll} and s_{sl} , respectively. These costs depend on the survey designs adopted. We assume a linear cost function. Hence, the total data-collection cost for the survey is given by:

$$C = c_0 + c_{ll}n_{ll} + c_{sl}n_{sl}. \quad (4)$$

Under this setup, the following result holds.

Theorem 1. *For an IST design that admits $\mathbb{V}(\hat{Y})$ in the form given by (3), the optimal sample size allocation under the linear cost function $C = c_0 + c_{ll}n_{ll} + c_{sl}n_{sl}$ is achieved by choosing*

$$n_{ll} = (C - c_0) \frac{\sqrt{A_z/c_{ll}}}{\sqrt{A_z c_{ll}} + \sqrt{A_t c_{sl}}}, \quad n_{sl} = (C - c_0) \frac{\sqrt{A_t/c_{sl}}}{\sqrt{A_z c_{ll}} + \sqrt{A_t c_{sl}}}. \quad (5)$$

The minimum variance of the estimator \hat{Y} is

$$\mathbb{V}(\hat{Y}) = \frac{1}{C - c_0} \left(\sqrt{A_z c_{ll}} + \sqrt{A_t c_{sl}} \right)^2 + B. \quad (6)$$

Proof. As in Särndal, Swensson, and Wretman (1992; Section 3.7.3), determining n_{ll} and n_{sl} to minimize $\mathbb{V}(\hat{Y})$ for fixed C is equivalent to minimizing the product

$$(\mathbb{V}(\hat{Y}) - B)(C - c_0) = \left(\frac{A_z}{n_{ll}} + \frac{A_t}{n_{sl}} \right) (c_{ll}n_{ll} + c_{sl}n_{sl}).$$

From the Cauchy–Schwarz inequality, we obtain:

$$(\mathbb{V}(\hat{Y}) - B)(C - c_0) \geq \left(\sqrt{A_z c_{ll}} + \sqrt{A_t c_{sl}} \right)^2,$$

where the equality holds if and only if:

$$\sqrt{\frac{c_{ll} n_{ll}}{A_z}} = \sqrt{\frac{c_{sl} n_{sl}}{A_t}} = K.$$

From the previous equality, it follows that

$$n_{ll} = K \sqrt{\frac{A_z}{c_{ll}}}, \quad n_{sl} = K \sqrt{\frac{A_t}{c_{sl}}}. \quad (7)$$

By replacing these quantities in the budget constraint (4), we obtain the value of K as:

$$K = \frac{c_{ll} n_{ll} + c_{sl} n_{sl}}{\sqrt{A_z c_{ll}} + \sqrt{A_t c_{sl}}} = \frac{C - c_0}{\sqrt{A_z c_{ll}} + \sqrt{A_t c_{sl}}},$$

which, when replaced in (7), yields (5). Hence, with this optimal choice of n_{ll} and n_{sl} , the quantity $(\mathbb{V}(\hat{Y}_{HT}) - B)(C - c_0)$ attains its minimum value $(\sqrt{A_z c_{ll}} + \sqrt{A_t c_{sl}})^2$ or, equivalently, $\mathbb{V}(\hat{Y})$ achieves the minimum variance bound given in (6). Hence the proof. \square

In terms of the sample size $n = n_{ll} + n_{sl}$, from (5) we have

$$n = (C - c_0) \frac{\sqrt{A_z/c_{ll}} + \sqrt{A_t/c_{sl}}}{\sqrt{A_z c_{ll}} + \sqrt{A_t c_{sl}}},$$

from which it easily follows that:

$$n_{ll} = n \frac{\sqrt{A_z/c_{ll}}}{\sqrt{A_z/c_{ll}} + \sqrt{A_t/c_{sl}}}, \quad n_{sl} = n \frac{\sqrt{A_t/c_{sl}}}{\sqrt{A_z/c_{ll}} + \sqrt{A_t/c_{sl}}}.$$

Hence, the following result is proved:

Corollary 1. *If the sample costs c_{ll} and c_{sl} are equal, the optimal sample size allocation is given by:*

$$n_{ll} = n \frac{\sqrt{A_z}}{\sqrt{A_z} + \sqrt{A_t}}, \quad n_{sl} = n \frac{\sqrt{A_t}}{\sqrt{A_z} + \sqrt{A_t}}. \quad (8)$$

We observe that the calculation of n_{ll} and n_{sl} given in (8) requires the knowledge of A_z and A_t . These quantities generally depend on the population variances that are usually unknown. When such values are unknown and cannot be properly guessed on the basis of previous data or experts opinion, they must be estimated making use, for instance, of a pilot survey (Sukhatme, Sukhatme, Sukhatme, & Asok, 1984).

4.2 | Allocation under simple random sampling without replacement

Let us suppose that the two samples s_{ll} and s_{sl} are selected according to simple random sampling without replacement (SRSWOR) and that all costs are equal. Hence, from (2), the variance of \hat{Y} can be reformulated as in (3):

$$\begin{aligned} \mathbb{V}(\hat{Y}) &= \mathbb{V}(\hat{Z}) + \mathbb{V}(\hat{T}) = \\ &= \left(1 - \frac{n_{ll}}{N}\right) \frac{S_z^2}{n_{ll}} + \left(1 - \frac{n_{sl}}{N}\right) \frac{S_t^2}{n_{sl}} = \\ &= \frac{S_z^2}{n_{ll}} + \frac{S_t^2}{n_{sl}} - \frac{1}{N} (S_z^2 + S_t^2), \end{aligned}$$

where S_z^2 denotes the population variance of the variable in the subscript. Note that $S_z^2 = S_y^2 + S_t^2 + 2S_{yt}$, where S_{yt} denotes the covariance. By replacing these population quantities by their sampling counterpart, we obtain an unbiased estimator of $\mathbb{V}(\hat{Y})$ as:

$$\hat{\mathbb{V}}(\hat{Y}) = \frac{s_z^2}{n_{ll}} + \frac{s_t^2}{n_{sl}} - \frac{1}{N}(s_z^2 + s_t^2)$$

where s^2 denotes the sample variance.

Finally, from (8), we have:

$$\gamma = \frac{n_{ll}}{n} = \frac{S_z}{S_z + S_t} = \frac{\sqrt{S_y^2 + S_t^2 + 2S_{yt}}}{\sqrt{S_y^2 + S_t^2 + 2S_{yt}} + S_t},$$

$$1 - \gamma = \frac{n_{sl}}{n} = \frac{S_t}{S_z + S_t} = \frac{S_t}{\sqrt{S_y^2 + S_t^2 + 2S_{yt}} + S_t}.$$

Clearly, if the correlation between the sensitive and the innocuous variables is positive, the LL-sample will be larger than the SL-sample. This is because the responses given in the LL-sample are expected to have a larger variance, which must be compensated with a larger sample size. Moreover, the function γ is: (i) an increasing function of S_y ; (ii) a decreasing function of S_t ; (iii) an increasing function of S_{yt} . Figure 1 shows the behavior of γ as a function of $S_y = 10, 20, \dots, 1000$ and $S_t = 10, 20, \dots, 1000$ for $\rho_{yt} = S_{yt}/S_y S_t = 0.5$.

4.3 | Allocation under a stratified sampling design

In the case of a stratified design, let the population U be divided into H strata. Let N_h denote the size of the h -th stratum, say U_h , and $W_h = N_h/N$ be the weight of U_h in the population, $h = 1, \dots, H$. From the stratum U_h , two samples $s_{h(ll)}$ and $s_{h(sl)}$ of sizes $n_{h(ll)}$ and $n_{h(sl)}$ are selected according to SRSWOR. The sampled elements in $s_{h(ll)}$ are confronted with the LL of items while those in $s_{h(sl)}$ are confronted with the SL of items. Under stratified SRSWOR, expression (2) takes the form:

$$\mathbb{V}(\hat{Y}_{str}) = \sum_{h=1}^H W_h^2 \left(1 - \frac{n_{h(ll)}}{N_h}\right) \frac{S_{h,z}^2}{n_{h(ll)}} + \sum_{h=1}^H W_h^2 \left(1 - \frac{n_{h(sl)}}{N_h}\right) \frac{S_{h,t}^2}{n_{h(sl)}}, \quad (9)$$

where $S_{h,z}^2$ is the variance in the stratum h .

As in Theorem 1, minimizing (9) subject to $\sum_{h=1}^H (n_{h(ll)} + n_{h(sl)}) = n$ with equal cost gives the following optimal sample size allocation for the stratum U_h :

$$n_{h(ll)} = n \frac{W_h S_{h,z}}{\sum_{h=1}^H W_h S_{h,z} + \sum_{h=1}^H W_h S_{h,t}}, \quad n_{h(sl)} = n \frac{W_h S_{h,t}}{\sum_{h=1}^H W_h S_{h,z} + \sum_{h=1}^H W_h S_{h,t}}.$$

Consequently:

$$\gamma_h = \frac{n_{h(ll)}}{n} = \sum_{h=1}^H W_h \frac{\sqrt{S_{h,y}^2 + S_{h,t}^2 + 2S_{h,yt}}}{\sum_{h=1}^H W_h \sqrt{S_{h,y}^2 + S_{h,t}^2 + 2S_{h,yt}} + \sum_{h=1}^H W_h S_{h,t}}$$

and

$$1 - \gamma_h = \frac{n_{h(sl)}}{n} = \sum_{h=1}^H W_h \frac{S_{h,t}}{\sum_{h=1}^H W_h \sqrt{S_{h,y}^2 + S_{h,t}^2 + 2S_{h,yt}} + \sum_{h=1}^H W_h S_{h,t}}.$$

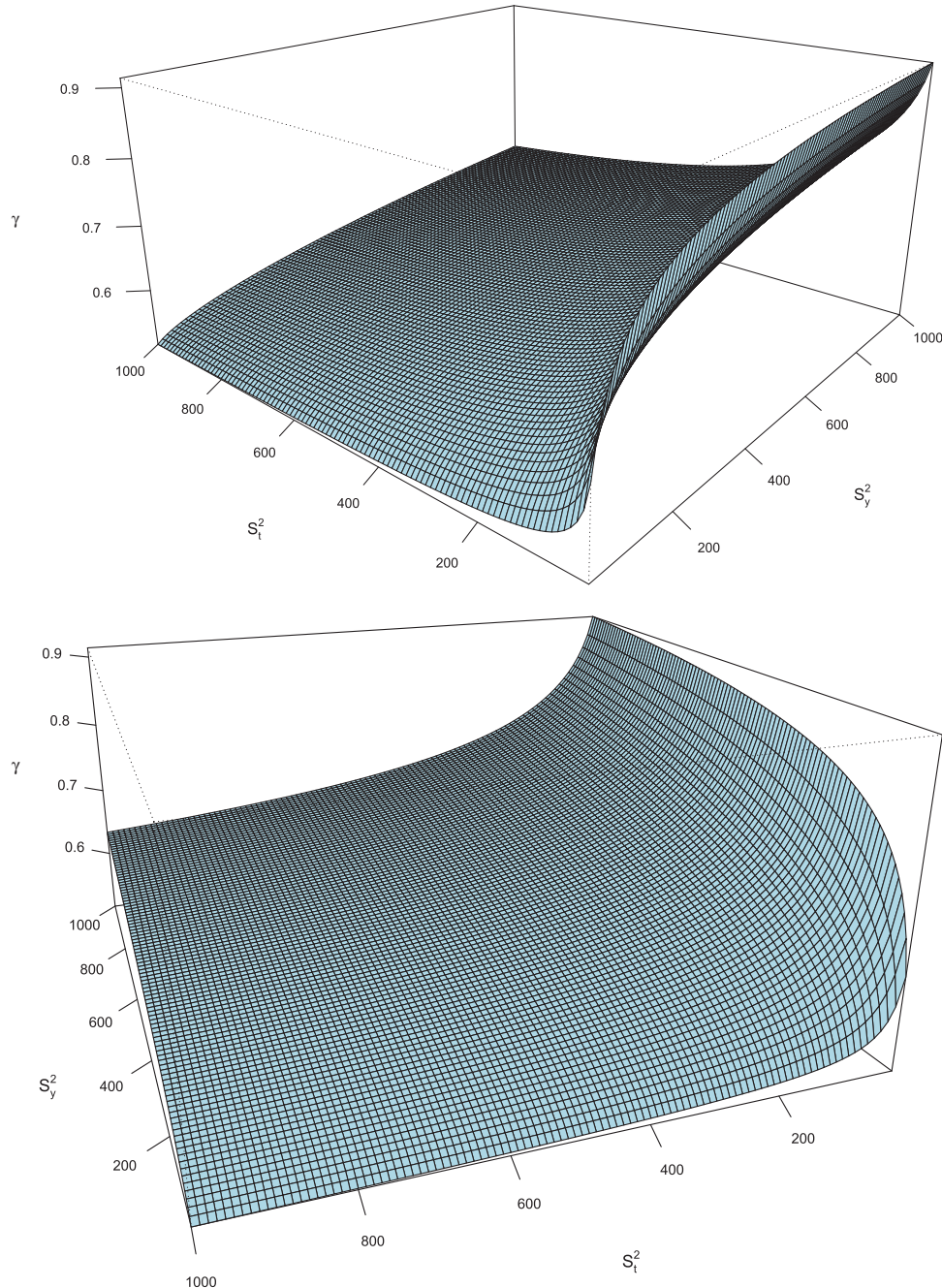


FIGURE 1 Behavior of γ for different values of S_y^2 and S_t^2 and $\rho_{yt} = 0.5$

4.4 | Allocation in multiple IST estimation

Determining optimal sample size allocation is of particular importance in the multiple IST estimation introduced in Section 3 where, under the separate and mixed approaches, more than two samples will be selected. Optimal allocation is easily achieved under the separate approach by applying the results of the previous sections to each sensitive variable under study. In other words, optimal sample size allocation is obtained for each IST survey by minimizing the variance of the estimator of the sensitive mean corresponding to the variable referred to by the IST survey. For the other approaches, the problem is slightly different but can be solved by extending the results of the previous sections after having specified the expression of the variance to be minimized. Let us first discuss the all-in-one procedure. In this case, just one sample is selected for the entire survey on the Q sensitive questions. This sample must then be optimally split into the LL-sample and SL-sample, and so the initial question is to decide how this optimality is to be achieved. One possibility is to focus on one of the Q sensitive variables, perhaps the most relevant variable—if

any—for the survey, and then to minimize the variance of the estimator of its mean. Obviously, however, obtaining the optimal sample size allocation for the variable considered does not ensure variance reduction in estimating the mean of the remaining variables. To overcome this limitation, a more general solution that involves all the study variables might be considered. Since multiple estimation leads to Q estimators of the Q population means of the sensitive variables under investigation, we may opt to minimize the variance of a convex combination of the Q variances of the estimators. Without loss of generality, let $\hat{Y}_k = \hat{Z}_k - \hat{T}_k$ denote the estimator of the population mean \bar{Y}_k for the sensitive variable \mathcal{Y}_k , $k = 1, \dots, Q$. The meaning of \hat{Z}_k and \hat{T}_k follows accordingly. Hence, the optimal sample sizes n_{ll} and n_{sl} are obtained by minimizing:

$$\mathbb{V}_\alpha = \sum_{k=1}^Q \alpha_k \mathbb{V}(\hat{Y}_k),$$

with $\sum_{k=1}^Q \alpha_k = 1$. For instance, under SRSWOR, for $Q = 2$ sensitive variables, say \mathcal{Y}_1 and \mathcal{Y}_2 , and $G = 2$ innocuous variables, say \mathcal{T}_1 and \mathcal{T}_2 , we have:

$$\mathbb{V}_\alpha = \frac{1}{n_{ll}} \left(\alpha_1 S_{z_1}^2 + \alpha_2 S_{z_2}^2 \right) + \frac{1}{n_{sl}} \left(\alpha_1 S_{t_1}^2 + \alpha_2 S_{t_2}^2 \right) - \frac{1}{N} \left[\alpha_1 \left(S_{z_1}^2 + S_{t_1}^2 \right) + \alpha_2 \left(S_{z_2}^2 + S_{t_2}^2 \right) \right]. \quad (10)$$

For the mixed approach, finding the optimal sample size allocation by minimizing the variance of one estimator is unfeasible since this will allocate the entire total size n between two samples, leaving a zero size for the remaining $Q - 1$ samples. The only solution to this problem is to minimize the convex combination of the Q variances of the estimators:

$$\begin{aligned} \mathbb{V}_\alpha &= \alpha_1 \mathbb{V}(\hat{Y}_1^*) + \alpha_2 \mathbb{V}(\hat{Y}_2^*) = \\ &= \alpha_1 \mathbb{V}(\hat{Z}_0) + \mathbb{V}(\hat{Z}_1) + \alpha_2 \mathbb{V}(\hat{Z}_2). \end{aligned}$$

For instance, if the samples are selected according to SRSWOR, $Q = 2$ and $G = 1$, we have:

$$\mathbb{V}_\alpha = \frac{1}{n_0} \alpha_1 S_t^2 + \frac{1}{n_1} S_{z_1}^2 + \frac{1}{n_2} \alpha_2 S_{z_2}^2 - \frac{1}{N} \left(\alpha_1 S_t^2 + S_{z_1}^2 + \alpha_2 S_{z_2}^2 \right). \quad (11)$$

Note that the choice $\alpha = 0.5$ is equivalent to minimizing $\mathbb{V}(\hat{Y}_1^*) + \mathbb{V}(\hat{Y}_2^*)$.

5 | SIMULATION

5.1 | Simulation design

In this section, we run a number of simulation studies to evaluate the performance of the optimal allocation discussed above. To do so, $N = 52,409$ artificial observations are generated for the sensitive variable \mathcal{Y} and the innocuous one \mathcal{T} . It is assumed that $(\mathcal{Y}, \mathcal{T})$ are observed from a bivariate normal distribution with different values of the correlation coefficient $\rho_{yt} = \rho$, and with mean and standard error vectors $\boldsymbol{\mu} = (3.114, 7.446)$ and $\boldsymbol{\sigma} = (0.604, 0.049)$, respectively. The values generated are then used to define the total score variable $\mathcal{Z} = \mathcal{Y} + \mathcal{T}$ and to obtain an estimate of \bar{Y} using: (i) the values of \mathcal{Y} in a standard HT-estimator as obtained by direct questioning; (ii) the values of \mathcal{T} and \mathcal{Z} in the HT-estimator as defined in (1). Hence, for each simulation study, we evaluate the estimated variance of the estimators for $B = 1000$ runs and for different sample sizes. Throughout the simulation, the costs are assumed to be constant.

The values for $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are taken from a real sensitive research conducted at the University of Granada in the academic year 2015/2016 to investigate the *consumption of cannabis*, using the IST. During the class time break, a sample of students were invited to participate in the study and to fill in a questionnaire. Some of these students (492) were directly posed the sensitive question: “How many cannabis cigarettes did you consume last year?” The remaining students (1293) were asked to provide data using the IST. In the IST survey, 773 students were arbitrarily allocated to the LL-sample and 520 to the SL-sample. The values $\mu = 3.114$ and $\sigma = 0.604$ represent the estimated mean and the estimated standard deviation for the number of cannabis cigarettes smoked. Similarly, the values $\mu = 7.446$ and $\sigma = 0.049$ refer to the estimates of the innocuous variable in the SL-sample. The innocuous variable \mathcal{T} is represented by students' score in the university entrance examinations (general stage score, ranging from 0 to 10). As a referee noted, the choice of this innocuous variable may not have sufficiently protected respondents'

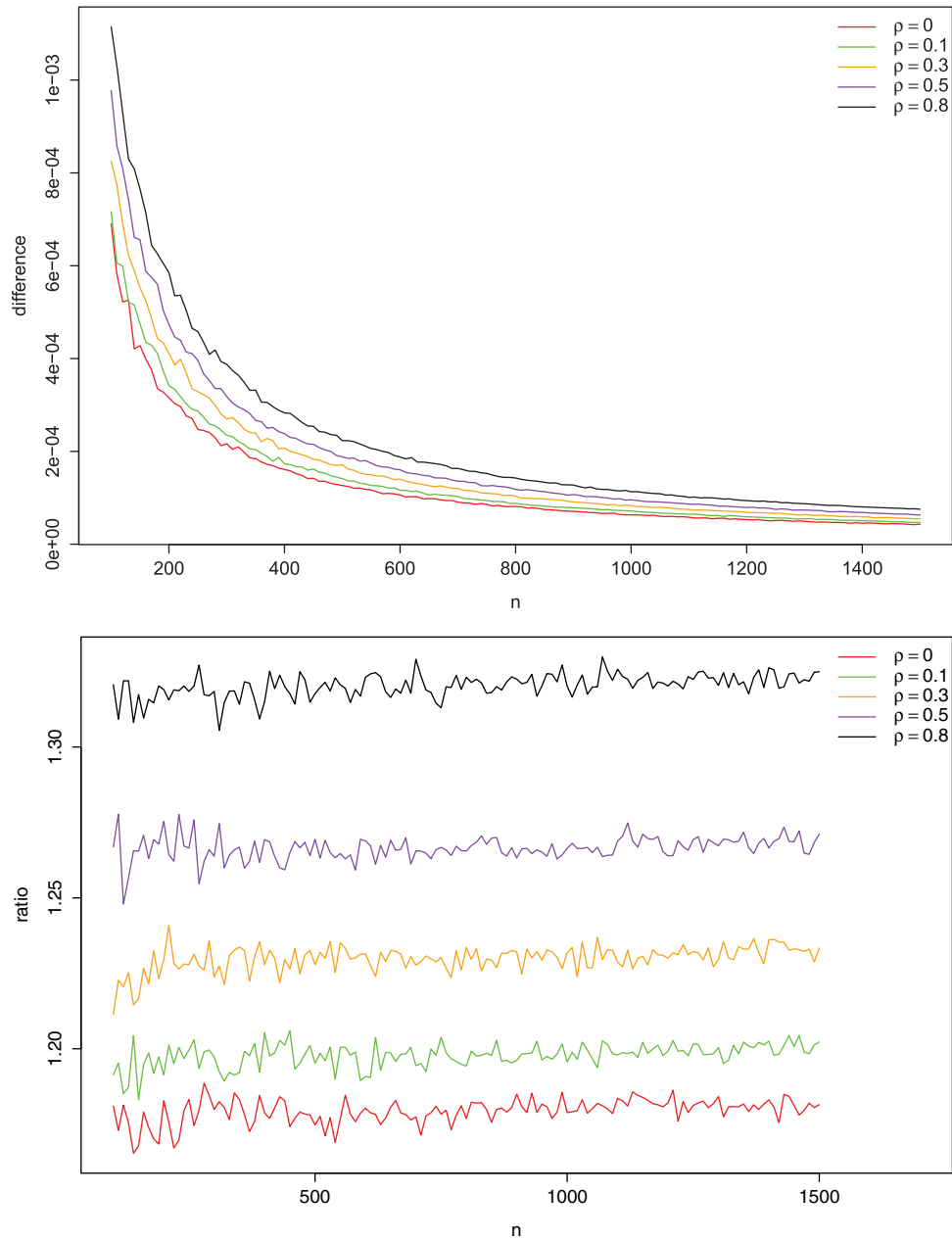


FIGURE 2 Difference and ratio between the variance of direct questioning estimates and optimal allocation IST estimates

privacy especially when the number of cannabis cigarettes smoked is “large,” for instance more than 50 cigarettes. Indeed, from the collected data, we observed that students who released IST responses (total scores) higher than 10 and 50 were 24.5% and 6.5%, respectively, and that nonresponse rate was very low (1.93%).

It is worthy noting that, according to the IST, 14.931 cannabis cigarettes were smoked on average, a value significantly higher than that obtained by direct questioning (one-tailed *t*-test, *p*-value < 0.001).

5.2 | Direct questioning versus optimal allocation IST estimates

In this first study, the samples are selected according to SRSWOR and the variance of the sample mean estimator $\bar{y} = \sum_{i \in S} y_i / n$ is compared with that of the IST estimator with optimal sample size allocation, performed on the same sample size *n*, as described in Section 4.2. Figure 2 illustrates the difference and the ratio between the estimated variances of the two estimators. Both the difference and the ratio are presented as mean values computed over *B* = 1000 replications. As expected, the variance of the IST estimator is higher than that of the sample mean estimator under direct questioning. The difference becomes negligible as the

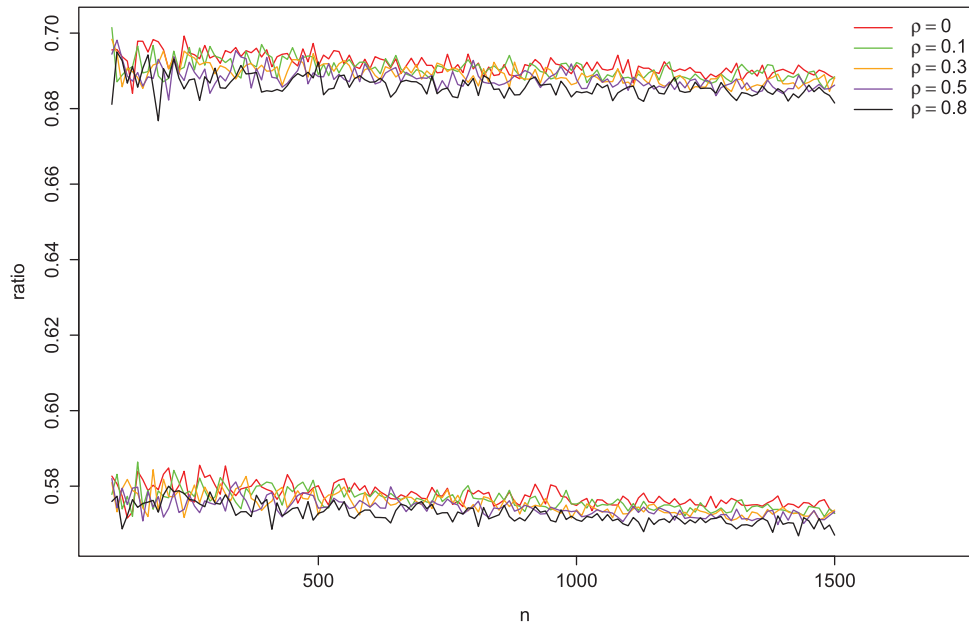


FIGURE 3 Ratio between the variance of the optimal allocation IST estimator and the variance of the IST estimator with $n_{II} = \lambda n$ under arbitrary allocation. The upper plots refers to $\lambda = 0.6$ and the lower to $\lambda = 0.5$

sample size increases, while the ratio highlights the fact that the loss of efficiency remains within acceptable limits especially when ρ is low. Moreover, for a fixed sample size, the difference (ratio) increases with ρ . The fact that the difference and the loss of efficiency are fairly modest values makes it clear that the optimal IST could provide estimates that are nearly as accurate as those obtained by direct questioning, and without jeopardizing respondents' confidentiality. This finding is of major importance in appraising the use of the IST in real surveys.

5.3 | Optimal versus arbitrary IST allocation

In SRSWOR, we now examine the efficiency gains that can be obtained when the IST allocation is optimal. To illustrate the magnitude of the increased efficiency, we consider the ratio between the variance of the optimal allocation IST estimator and that of the IST estimator arbitrarily obtained assuming $n_{II} = \lambda n$ and $n_{sI} = (1 - \lambda)n$, $\lambda = 0.5, 0.6$. The results are shown in Figure 3. The improved efficiency is evident in both situations. As also shown in Figure 2, the correlation coefficient does not appear to significantly affect the variance of the IST estimators and, consequently, the efficiency gain from the optimal allocation.

5.4 | Optimal IST allocation in stratified SRSWOR

We now examine the case in which stratified SRSWOR is adopted. We assume that the $N = 52,409$ students of the University of Granada (see Section 5.1) are stratified into two groups—male (M) and female (F)—with weights $W_M = 0.442$ and $W_F = 0.558$ known from administrative sources. Under the same framework as in Section 5.1, for the male group we generate $N_M = 23,151$ observations from the bivariate normal distribution $(\mathcal{Y}, \mathcal{T})$ with different values of ρ , $\mu_M = (6.340, 7.507)$ and $\sigma_M = (1.431, 0.072)$. Similarly, for the female stratum ($N_F = 29,258$), we assume $\mu_F = (0.240, 7.408)$ and $\sigma_F = (0.121, 0.067)$. As in Section 5.1, the entries of the vectors μ and σ represent the estimated means and the estimated standard deviations of the sensitive variable and the innocuous variable computed from the male/female direct questioning samples and for the male/female SL-samples, respectively.

The minimum variance estimator of the stratified IST estimator is achieved by using the optimal sample size allocation given in Section 4.3. The variance of the estimates under optimal allocation is then compared using two different forms of allocation:

- (i) **Arbitrary allocation.** In stratified IST with two strata (U_M and U_F), four samples are considered. From the U_M stratum, the LL-sample and the SL-sample are selected. Similarly, for the U_F stratum. Let n_{II} and n_{sI} be the sample sizes in the respective groups. Hence, we trivially assume: $n_{II|M} = n_{sI|M} = n_{II|F} = n_{sI|F} = n/4$.

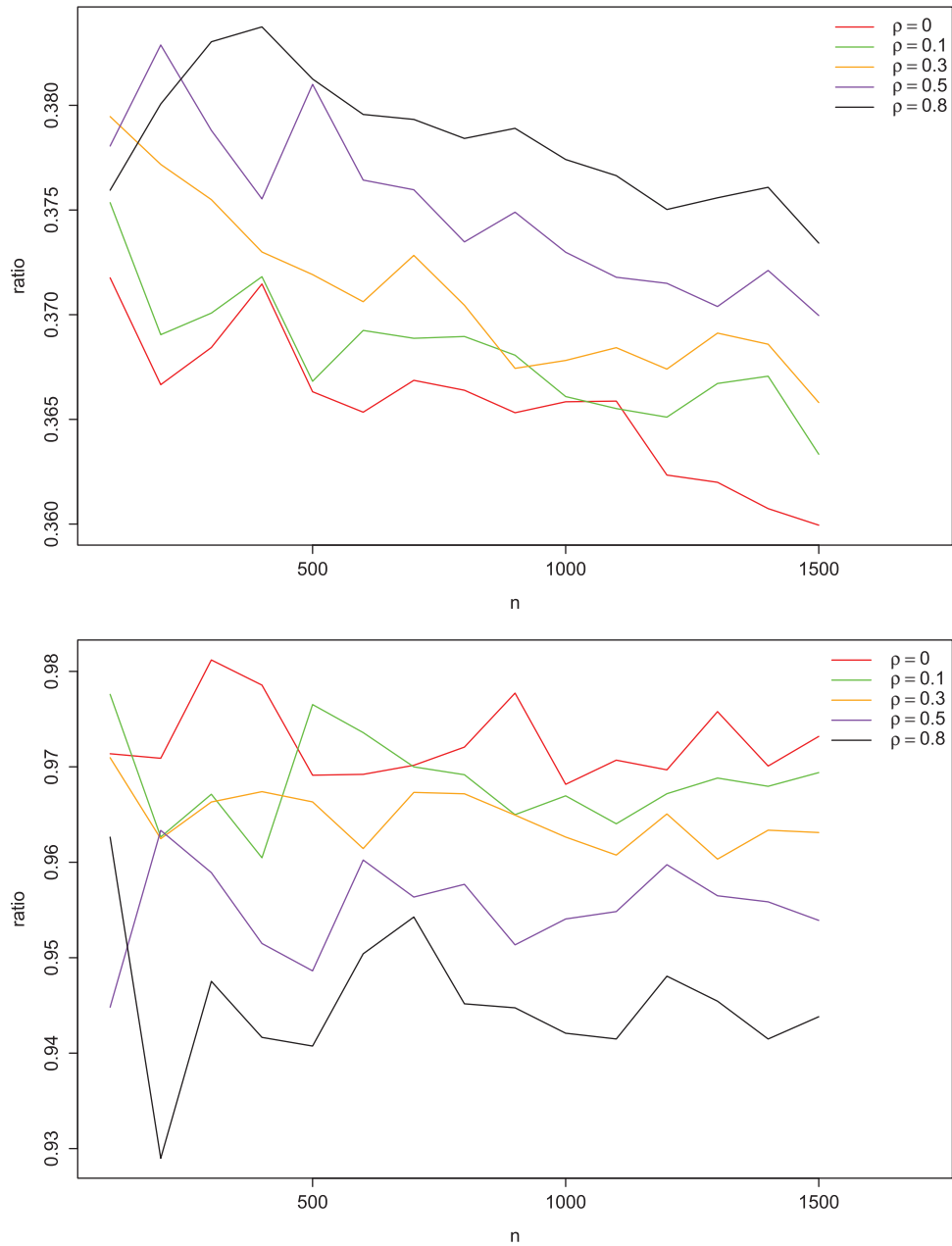


FIGURE 4 Ratio between variance under optimal allocation and under: (i) arbitrary allocation (upper plot), (ii) naive two-step optimal allocation (lower plot)

(ii) Naive two-step optimal allocation. Allocation is conducted in two steps, separately determining the optimal IST allocation in one sample of men and in another of women. In the first step, a stratified sample of male and female students is selected with proportional allocation (see, e.g. Särndal et al., 1992). In the second step, each of the two first-step samples is optimally allocated in the LL-sample and SL-sample according to (8).

Figure 4 shows the ratio between the variances of the optimal and non-optimal allocation stratified IST estimators. It can be seen that arbitrary allocation is not at all efficient, while the results obtained with two-step allocation are almost identical to those attainable with the theoretical optimal allocation.

Finally, we compared the efficiency of stratified and SRSWOR IST estimates under optimal allocation. The results shown in Figure 5 reflect the considerable gain in efficiency achieved by stratifying the population.

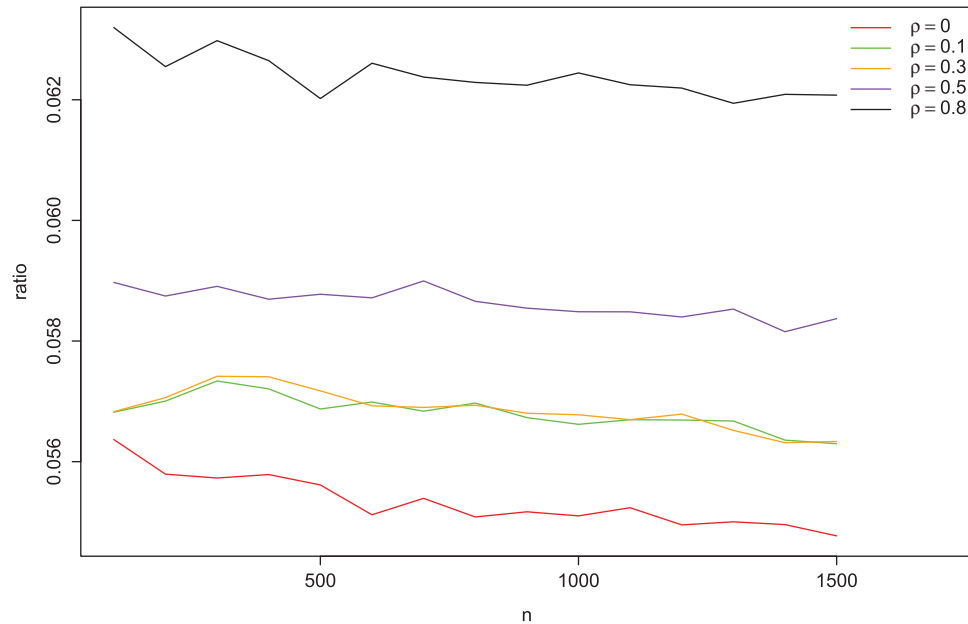


FIGURE 5 Ratio of optimal allocation variances under stratified IST and SRSWOR IST

5.5 | Optimal allocation in multiple IST estimation

In this section, we investigate multiple IST estimation under each of the approaches discussed in Section 3. The simulation study is based on real data from the *Survey of Household Income and Wealth* (SHIW) conducted by the Bank of Italy (2014). The survey covers 8156 households composed of 19,366 individuals. We assume the 8156 households as the target population and focus on two sensitive variables: (i) *net disposable income* (\mathcal{Y}_1), and (ii) *net wealth* (\mathcal{Y}_2). For all the households surveyed, the values of these and other variables are known.

The aim of this simulation study is to compare the IST estimates of \bar{Y}_1 and \bar{Y}_2 under the separate, all-in-one and mixed approaches by assuming that $\mathcal{T} = \text{consumption}$ is the innocuous variable for implementing the IST. From the available data, we know that $\bar{Y}_1 = 31,248$ euro, $\bar{Y}_2 = 236,097$ euro and these values are used as benchmarks. Under the separate approach, the optimal sample allocation for n_{I1} and n_{s1} is separately considered for each of the two variables in such a way that the estimates for \bar{Y}_1 and \bar{Y}_2 both attain their minimum variance bound. The all-in-one estimates are obtained assuming that data on both the variables are collected by performing the IST twice on the same units belonging to the only sample selected. The optimal sample sizes n_{I1} and n_{s1} , which minimize (10) with $\alpha = 0.5$, are used to obtain the estimates of \bar{Y}_1 and \bar{Y}_2 . Obviously, using n_{I1} and n_{s1} does not ensure that minimum variance is achieved for \hat{Y}_1 and \hat{Y}_2 . A similar procedure is employed for the mixed approach. In this case, the three sample sizes n_0 , n_1 and n_2 are optimally determined to minimize (11) with $\alpha = 0.5$ and then used in the single estimators \hat{Y}_1^* and \hat{Y}_2^* . We specify that in all situations the optimal allocation has been achieved by minimizing the estimated variance.

For different sample sizes and $B = 1000$ replications, we investigate the performance of the estimators under the three approaches by means of the absolute relative bias (ARB) and the relative variance (RV):

$$\text{ARB}(\hat{\theta}_i) = \frac{\sum_{k=1}^B |\hat{\theta}_i^{(k)} - \bar{Y}_i|}{B\bar{Y}_i}, \quad \text{RV}(\hat{\theta}_i) = \frac{\sum_{k=1}^B (\hat{\theta}_i^{(k)} - \bar{Y}_i)^2}{B\bar{Y}_i^2}$$

with $\hat{\theta}_i^{(k)}$ denoting the estimate of \bar{Y}_i on the k -th sample selected from the SHIW target population according to SRSWOR.

The outcomes of the simulation are summarized in Figure 6. It is immediately apparent that both the ARB and the RV decrease as the sample size increases, which is a clear indication of the consistency of the estimates under the three approaches. The three approaches produce equivalent results in estimating the mean of $\mathcal{Y}_2 = \text{wealth}$, while for $\mathcal{Y}_1 = \text{income}$ the separate approach seems to slightly outperform the others, especially for usual sample sizes. As the sample size increases, the difference between the methods decreases. However, on the whole there are no striking differences and for the situations considered in this analysis, the mixed approach seems to be competitive in terms of efficiency while clearly reducing the statistical burden

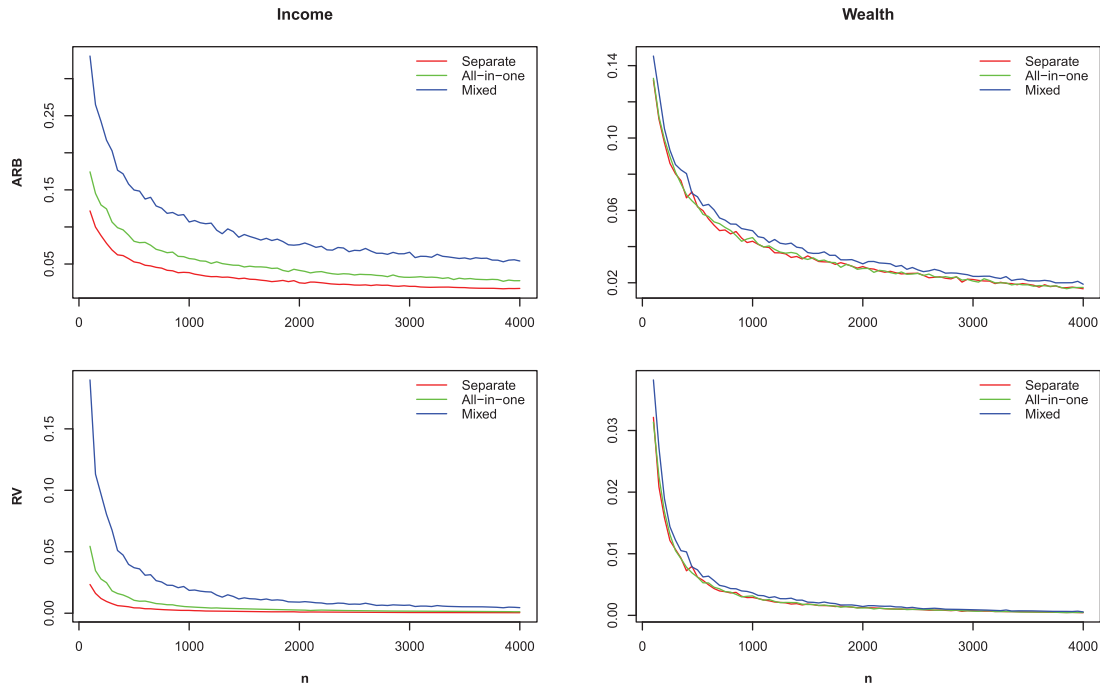


FIGURE 6 Performance of the estimates under the three IST approaches for multiple estimates purposes. The results are based on Monte Carlo simulated ARB and RV

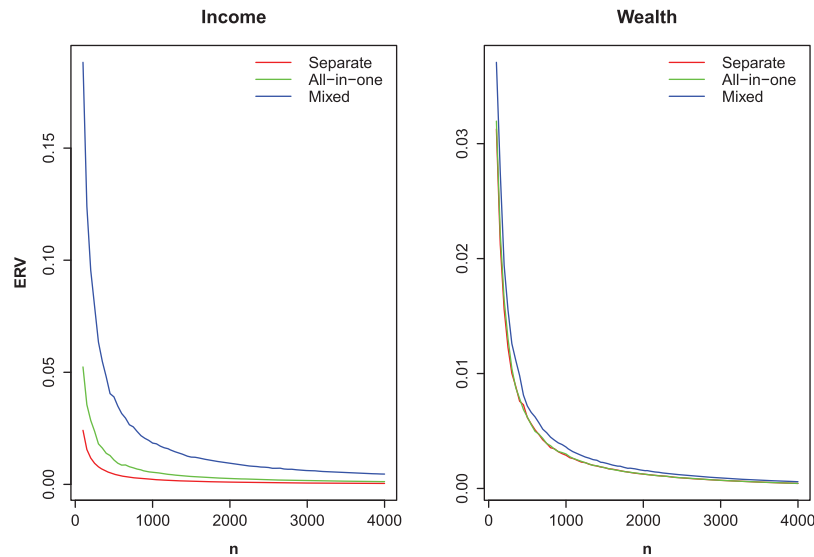


FIGURE 7 Performance of the estimates under the three IST approaches for multiple estimates purposes. The results are based on estimated theoretical variances

on the respondents. We then replicated the simulation study by directly comparing the theoretical estimated variances of the estimators of \bar{Y}_1 and \bar{Y}_2 under the three approaches. Figure 7 shows the behavior of the estimated relative variance (ERV), obtained by dividing the estimated variance of $\hat{\theta}_i$ by \bar{Y}_i^2 , $i = 1, 2$. The results obtained confirm those for the RV reported in Figure 6. We conclude, therefore, that multiple estimation may be profitably pursued via different approaches and that a useful trade-off between efficiency in the estimates and reducing the statistical burden may be achieved by using the mixed approach with optimal allocation. The findings of this study may therefore be of major significance to survey statisticians and practitioners to support the use of the IST in real-world studies.

6 | CONCLUSIONS

The IST enables us to estimate the mean (or the total) of stigmatizing quantitative variables using an indirect questioning approach, thus reducing nonresponse rates and social desirability response bias. This method is closely related to the ICT, which was developed to measure the proportion of dichotomous sensitive items in human population surveys.

In this article, we presented certain methodological advances in the use of the IST, and discussed two open questions. First, we considered the problem of how to reduce the statistical burden on respondents when $Q \geq 2$ sensitive variables are surveyed and the population means need to be estimated. Three ways of applying the IST have been discussed. The first of these, the separate approach, requires that for each sensitive item one LL-sample and one SL-sample be selected, that is in total, $2Q$ samples are used. In the second approach, termed all-in-one, one LL-sample, and one SL-sample are selected and the respondents are asked to participate in Q distinct IST surveys, one for each sensitive item. The separate approach is time-consuming and costly, while the all-in-one approach places an excessive burden on the survey participants that could even induce them to break the rules of the IST or to drop out of the survey. Given the weaknesses of these two approaches, a viable alternative providing a possible trade-off could be pursued. A mixed approach, which requires the selection of $Q + 1$ independent samples, has been therefore proposed, and its performance investigated through a number of simulation experiments based on optimal sample size allocation.

The optimal allocation of the total sample size into the LL-sample and the SL-sample is the second, but no less important, issue discussed in this article. First, we considered a method of allocation based on minimizing the variance of the IST estimator of the mean of one sensitive variable that is valid under a budget constraint and for a general sampling design. Thus, explicit expressions for the sampling fractions have been worked out when the SRSWOR and stratified sampling designs are used. The allocation method has been then extended to the case of Q sensitive variables under the all-in-one and mixed approaches.

An extensive simulation study has been conducted to investigate the performance of the proposed techniques and the related estimators under different sampling designs and for different sample sizes. All the situations examined reflect the benefits of determining the optimal sample size, which can significantly increase the efficiency of the estimates with respect to any arbitrary allocation of the sample units.

A very interesting result has been achieved when optimal allocation is used for multiple IST estimation purposes under the mixed approach. In this case, in relation to the marked reduction obtained in the statistical burden placed on respondents and in survey costs, the loss of efficiency with respect to the all-in-one and separate approaches may be considered very modest or even negligible. Hence, from a theoretical standpoint, the mixed approach appears to be a viable alternative for the purposes of multiple IST estimation. That said, final users interested in experiencing multiple IST have enough elements to critically evaluate the feasibility of the different procedures and to weight between pros and cons with regards to costs, time effort, respondents' burden, and accuracy.

We conclude by observing that all the ideas, the methodological advances and the results presented in this article regarding the IST may be easily extended to its forerunner, the ICT, which, although it is a more widespread and long-established technique, suffers from the same drawbacks that are discussed in this article with respect to the IST and that, to our knowledge, have not yet been addressed. Hence, the value of this article is twofold.


ACKNOWLEDGMENT

This work is partially supported by Ministerio de Economía y Competitividad (grant MTM2015-63609-R, Spain), Ministerio de Educación, Cultura y Deporte (grant FPU, Spain), and by the project PRIN-SURWEY (grant 2012F42NS8, Italy).

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

ORCID

Pier Francesco Perri  <http://orcid.org/0000-0002-3432-800X>

REFERENCES

- Arcos, A., Rueda, M., & Singh, S. (2015). Generalized approach to randomized response for quantitative variables. *Quality and Quantity*, *49*, 1239–1256.
- Arentoft, A., Van Dyk, K., Thames, A. D., Sayegh, P., Thaler, N., Schonfeld, D., .. Hinkin, C. H. (2016). Comparing the unmatched count technique and direct self-report for sensitive health-risk behaviors in HIV+ adults. *AIDS Care*, *28*, 370–375.
- Arias, A., & Sutton, S. G. (2013). Understanding recreational fishers compliance with no-take zones in the Great Barrier Reef Marine Park. *Ecology and Society*, *18*. Available at <https://doi.org/10.5751/ES-05872-180418>.
- Arnab, R., & Singh, S. (2010). Randomized response techniques: An application to the Botswana AIDS impact survey. *Journal of Statistical Planning and Inference*, *140*, 941–953.
- Bank of Italy. (2014). Survey of household income and wealth. Available at <https://www.bancaditalia.it/statistiche/tematiche/indagine-famiglie-imprese/bilanci-famiglie/index.html?com.dotmarketing.htmlpage.language=1>.
- Barabesi, L., Diana, G., & Perri, P. F. (2013). Design-based distribution function estimation for stigmatized populations. *Metrika*, *76*, 919–935.
- Barabesi, L., Diana, G., & Perri, P. F. (2015). Gini index estimation in randomized response surveys. *Asta-Advances in Statistical Analysis*, *99*, 45–62.
- Blank, S. G., & Gavin, M. C. (2009). The randomized response technique as a tool for estimating non-compliance rates in fisheries: A case study of illegal red abalone (*Haliotis rufescens*) fishing in Northern California. *Environmental Conservation*, *36*, 112–119.
- Chaudhuri, A. (2011). *Randomized response and indirect questioning techniques in surveys*. Chapman & Hall/CRC, Boca Raton, FL.
- Chaudhuri, A., & Christofides, T. C. (2013). *Indirect questioning in sample surveys*. Springer-Verlag, Berlin, Heidelberg, DE.
- Chaudhuri, A., & Mukerjee, R. (1988). *Randomized response: Theory and techniques*. Marcel Dekker, Inc., New York, NY.
- Chen, X., Du, Q., Jin, Z., Xu, T., Shi, J., & Gao, G. (2014). The randomized response technique application in the survey of homosexual commercial sex among men in Beijing. *Iran Journal of Public Health*, *43*, 416–422.
- Cobo, B., Rueda, M. D. M., & López-Torrecillas, F. (2016). Application of randomized response techniques for investigating cannabis use by Spanish university students. *International Journal of Methods in Psychiatric Research*, first published online, <https://doi.org/10.1002/mpr.1517>.
- Conteh, A., Gavin, M. C., & Solomon, J. (2015). Quantifying illegal hunting: A novel application of the randomised response technique. *Biological Conservation*, *189*, 16–23.
- Cross, P., Edwards-Jones, G., Omed, H., & Williams, A. P. (2010). Use of a randomized response technique to obtain sensitive information on animal disease prevalence. *Preventive Veterinary Medicine*, *96*, 252–262.
- De Jong, M. G., Pieters, R., & Stremersch, S. (2012). Analysis of sensitive questions across cultures: An application of multigroup item randomized response theory to sexual attitudes and behavior. *Journal of Personality and Social Psychology*, *103*, 543–564.
- Diana, G., & Perri, P. F. (2011). A class of estimators for quantitative sensitive data. *Statistical Papers*, *52*, 633–650.
- Dickson-Swift, V., James, E. L., & Liamputtong, P. (2008). *Undertaking sensitive research in the health and social sciences. Managing boundaries, emotions and risks*. Cambridge University Press, New York, NY.
- Dietz, P., Striegel, H., Franke, A. G., Lieb, K., Simon, P., & Ulrich, R. (2013). Randomized response estimates for the 12-month prevalence of cognitive-enhancing drug use in university students. *Pharmacotherapy*, *33*, 44–50.
- Fox, J. A., & Tracy, P. E. (1986). *Randomized response: A method for sensitive survey*. Sage Publication, Inc., Newbury Park, CA.
- Fox, J. P., Avetisyan, M., & van der Palen, J. (2013). Mixture randomized item-response modeling: A smoking behavior validation study. *Statistics in Medicine*, *32*, 4821–4837.
- Fox, J. P., Entink, R. K., & Avetisyan, M. (2014). Compensatory and non-compensatory multidimensional randomized item response models. *British Journal of Mathematical and Statistical Psychology*, *67*, 133–152.
- Franke, A. G., Bagusat, C., Dietz, P., Hoffmann, I., Simon, P., Ulrich, R., & Lieb, K. (2013). Use of illicit and prescription drugs for cognitive or mood enhancement among surgeons. *BMC Medicine*, *11*, 102.
- Geng, G. Z., Gao, G., Ruan, Y. H., Yu, M. R., & Zhou, Y. H. (2016). Behavioral risk profile of men who have sex with men in Beijing, China: Results from a cross-sectional survey with randomized response techniques. *Chinese Medical Journal*, *129*, 523–529.
- Glynn, A. N. (2013). What can we learn with statistical truth serum? Design and analysis of the list experiment. *Public Opinion Quarterly*, *77*, 159–172.
- Groenitz, H. (2014). Improvements and extensions of the item count technique. *Electronic Journal of Statistics*, *8*, 2321–2351.
- Gunaratne, A., Kubota, S., Kumarawadu, P., Karunagoda, K., & Kon, H. (2016). Is hiding foot and mouth disease sensitive behavior for farmers? A survey study in Sri Lanka. *Asian-Australasian Journal of Animal Sciences*, *29*, 280–287.
- Harrison, M., Baker, J., Twinamatsiko, M., & Milner-Gulland, E. J. (2015). Profiling unauthorized natural resource users for better targeting of conservation interventions. *Conservation Biology*, *29*, 1636–1646.
- Hoffmann, A., & Musch, J. (2016). Assessing the validity of two indirect questioning techniques: A stochastic lie detector versus the crosswise model. *Behavior Research Methods*, *48*, 1032–1046.
- Hoffmann, A., Diedenhofen, B., Verschuere, B., & Musch, J. (2016). A strong validation of the crosswise model using experimentally-induced cheating behavior. *Experimental Psychology*, *62*, 403–414.

- Hussain, Z. Shabbir, N., & Shabbir J., (2015). An alternative item sum technique for improved estimators of population mean in sensitive surveys. *Hacettepe Journal of Mathematics and Statistics*, first published online, <https://doi.org/10.15672/HJMS.20159113160>.
- Ibrahim, F. (2016). An alternative modified item count technique in sampling survey. *International Journal of Statistics and Applications*, 6, 177–187.
- Imai, K. (2011). Multivariate regression analysis for the item count technique. *Journal of the American Statistical Association*, 106, 407–416.
- Imai, K., Park, B., & Greene, K. F. (2015). Using the predicted responses from list experiments as explanatory variables in regression models. *Political Analysis*, 23, 180–196.
- James, R. A., Nepusz, T., Naughton, D. P., & Petróczi, A. (2013). A potential inflating effect in estimation models: Cautionary evidence from comparing performance enhancing drug and herbal hormonal supplement use estimates. *Psychology of Sports and Exercise*, 14, 84–96.
- Kazemzadeh, Y., Shokoohi, M., Baneshi, M. R., & Haghdoost, A. A. (2016). The frequency of high-risk behaviors among Iranian college students using indirect methods: Network scale-up and crosswise model. *International Journal of High Risk Behaviors & Addictions*, first published online, <https://doi.org/10.5812/ijhrba.25130>.
- Khosravi, A., Mousavi, S. A., Chaman, R., Khosravi, F., Amiri, M., & Shamsipour, M. (2015). Crosswise model to assess sensitive issues: A study on prevalence of drug abuse among university students of Iran. *International Journal of High Risk Behaviors & Addictions*, first published online, <https://doi.org/10.5812/ijhrba.24388v2>
- Krebs, C. P., Lindquist, C. H., Warner, T. D., Fisher, B. S., Martin, S. L., & Childers, J. M. (2011). Comparing sexual assault prevalence estimates obtained with direct and indirect questioning techniques. *Violence Against Women*, 17, 219–235.
- Liu, Y., & Tian, G. L. (2013). Multi-category parallel models in the design of surveys with sensitive questions. *Statistics and Its Interface*, 6, 137–142.
- Miller, J. D. (1984). *A new survey technique for studying deviant behavior*. Ph.D. thesis. The George Washington University, Washington, DC.
- Moshagen, M., Hilbig, B. E., Erdfelder, E., & Moritz, A. (2014). An experimental validation method for questioning techniques that assess sensitive issues. *Experimental Psychology*, 61, 48–54.
- Moshagen, M., Musch, J., Ostapczuk, M., & Zhao, Z. (2010). Reducing socially desirable responses in epidemiologic surveys: An extension of the randomized-response technique. *Epidemiology*, 21, 379–382.
- Moseson, H., Massaquoi, M., Dehlendorf, C., Bawo, L., Dahn, B., Zolia, Y., .. Gerdtts, C. (2015). Reducing under-reporting of stigmatized health events using the list experiment: Results from a randomized, population-based study of abortion in Liberia. *International Journal of Epidemiology*, 44, 1951–1958.
- Nakhaee, M. R., Pakravan, F., & Nakhaee, N. (2013). Prevalence of use of anabolic steroids by bodybuilders using three methods in a city of Iran. *Addiction and Health*, 5, 77–82.
- Nepusz, T., Petróczi, A., Naughton, D. P., Epton, T., & Norman, P. (2014). Estimating the prevalence of socially sensitive behaviors: Attributing guilty and innocent noncompliance with the single sample count method. *Psychological Methods*, 19, 334–355.
- Nuno, A., Bunnefeld, N., Naiman, L. C., & Milner-Gulland, E. J. (2013). A novel approach to assessing the prevalence and drivers of illegal bushmeat hunting in the Serengeti. *Conservation Biology*, 2, 1355–1365.
- Oliveras, E., & Letamo, G. (2010). Examples of methods to address underreporting of induced abortion: Preceding birth technique and randomized response technique. In Singh S., Remez L., Tartaglione A. (Eds.), *Methodologies for estimating abortion incidence and abortion-related morbidity: A review*. Guttmacher Institute, New York - International Union for the Scientific Study of the Population, Paris.
- Perri, P. F., & van der Heijden, P. G. M. (2012). A property of the CHAID partitioning method for dichotomous randomized response data and categorical predictors. *Journal of Classification*, 29, 76–90.
- Perri, P. F., Pelle, E., & Stranges, M. (2016). Estimating induced abortion and foreign irregular presence using the randomized response crossed model. *Social Indicators Research*, 129, 601–618.
- Petróczi, A., Nepusz, T., Cross, P., Taft, H., Shah, S., Deshmukh, N., .. Naughton, D. P. (2011). New non-randomised model to assess the prevalence of discriminating behaviour: A pilot study on mephedrone. *Substance Abuse Treatment, Prevention, and Policy*, 6, 20.
- Randrianantoandro, T. N., Kono, H., & Kubota, S. (2015). Knowledge and behavior in an animal disease outbreak: Evidence from the item count technique in a case of African swine fever in Madagascar. *Preventive Veterinary Medicine*, 118, 483–487.
- Rueda, M., Cobo, B., & Arcos, A. (2016). An improved class of estimators in RR surveys. *Mathematical Methods in the Applied Sciences*, first published online, <https://doi.org/10.1002/mma.4256>.
- Safiri, S. (2016). Knowledge, attitude, self-efficacy and estimation of frequency of condom use among Iranian students based on a crosswise model: More explanation is needed for the crosswise model. *International Journal of Adolescent Medicine and Health*, first published online, <https://doi.org/10.1515/ijamh-2016-0110>.
- Särndal, C. E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. Springer-Verlag, New York, NY.
- Shamsipour, M., Yunesian, M., Fotouhi, A., Jann, B., Rahimi-Movaghar, A., Asghari, F., & Akhlaghi, A. A. (2014). Estimating the prevalence of illicit drug use among students using the crosswise model. *Substance Use and Misuse*, 49, 1303–1310.
- Starosta, A. J., & Earleywine, M. (2014). Assessing base rates of sexual behavior using the unmatched count technique. *Health Psychology and Behavioral Medicine*, 2, 198–210.

- Striegel, H., Ulrich, R., & Simon, P. (2010). Randomized response estimates for doping and illicit drug use in elite athletes. *Drug and Alcohol Dependence*, *106*, 230–232.
- Stubbe, J. H., Chorus, A. M. J., Frank, L. E., de Hon, O., & van der Heijden, P. G. M. (2013). Prevalence of use of performance enhancing drugs by fitness center members. *Drug Testing and Analysis*, *6*, 434–438.
- Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S., & Asok, C. (1984) *Sampling theory of survey with applications*. Iowa State University Press, Ames, IA.
- Tian, G. L., & Tang, M. L. (2014). *Incomplete categorical data design: Non-randomized response techniques for sensitive questions in surveys*. Chapman & Hall/CRC, Boca Raton, FL.
- Trappmann, M., Krumpal, I., Kirchner, A., & Jann, B. (2014). Item sum: A new technique for asking quantitative sensitive questions. *Journal of Survey Statistics and Methodology*, *2*, 58–77.
- Tsuchiya, T. (2005). Domain estimators for the item count technique. *Survey Methodology*, *31*, 41–51.
- Tu, S. H., & Hsieh, S. H. (2017). Estimates of lifetime extradyadic sex using a hybrid of randomized response technique and crosswise design. *Archives of Sexual Behavior*, *46*, 373–384.
- Ulrich, R., Schörter, H., Striegel, H., & Simon, P. (2012). Asking sensitive questions: A statistical power analysis of randomized response models. *Psychological Methods*, *17*, 623–641.
- Vakilian, K., Mousavi, S. A., Keramat, A., & Chaman, R. (2016). Knowledge, attitude, self-efficacy and estimation of frequency of condom use among Iranian students based on a crosswise model. *International Journal of Adolescent Medicine and Health*, first published online, <https://doi.org/10.1515/ijamh-2016-0010>.
- Wu, Q., & Tang, M. L. (2016). Non-randomized response model for sensitive survey with noncompliance. *Statistical Methods in Medical Research*, *25*, 2827–2839.

SUPPORTING INFORMATION

Additional Supporting Information including source code to reproduce the results may be found online in the supporting information tab for this article.

How to cite this article: Perri PF, Rueda García MdM, Cobo Rodríguez B. Multiple sensitive estimation and optimal sample size allocation in the item sum technique. *Biometrical Journal*. 2018;60:155–173. <https://doi.org/10.1002/bimj.201700021>