Preserving privacy protection using indirect questioning techniques in real sensitive surveys

Pier Francesco Perri (pierfrancesco.perri@unical.it)[[1]](#footnote-1)

**Keywords:** Item count technique, Item sum technique, Multiple frames surveys, Randomized response theory*,* Social desirability bias

# Introduction

Nowadays, large scale surveys are increasingly delving into sensitive topics such as religious prejudice, racism, drug use, sexual behaviour, gambling, consumption of alcohol, domestic violence. Sensitive, stigmatizing or even incriminating themes are difficult to investigate by means of standard data-collection survey techniques since respondents are generally reluctant to release information concerning their personal sphere. Consequently, doing research on delicate topics is not an easy matter since it is likely to meet with three sources of errors: (1) refusal to cooperate (unit-non-response); (2) refusal to answer specific questions (item-non-response); (3) untruthful answers (measurement error). In particular, dishonest or misleading answers generate a well-known source of bias which is called *social desirability bias*, i.e. the tendency of survey participants to present themselves in a positive light. All these errors can seriously flaw the quality of the data and, thus, jeopardize the usefulness of the collected information for subsequent analyses, including inference on unknown characteristics of the population under study. More specifically, standard survey questioning techniques based on self-reporting or direct questions generally produce overreporting of socially acceptable attitudes which conform to social norms and underreporting of socially disapproved, undesirable behaviours which deviate from social rules.

To reduce the problem, the level of cooperation obtained from survey participants must be increased. Since the decision to cooperate, in terms of providing complete and honest answers, depends on how respondents perceive their privacy will be protected, survey modes which ensure full anonymity go some way to increasing the chance of cooperation and, consequently, that of obtaining more reliable information on sensitive issues. In this respect, survey statisticians have developed many different strategies to ensure respondents’ anonymity and to reduce the incidence of evasive answers and misreporting of social taboos when direct questions are posed on sensitive themes. One possibility for improving reporting is to limit the influence of the interviewer in the question and answer process, as the presence of the interviewer tends to increase socially desirability effects. This goal is traditionally pursued by means of self-administered questionnaires with paper and pencil, the interactive voice response technique, computer-assisted telephone interviewing, computer-assisted self interviewing, audio computer-assisted self interviewing or by computer-assisted Web interviewing. Alternatively, since the 1960s a variety of questioning methods have been devised to ensure respondents' anonymity. These methods are generally known as *indirect questioning techniques* (IQTs; for a review see Chaudhuri and Christofides [1]) and they obey 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 involved in sensitive behaviours. In this way, privacy is protected since responses 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 data 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 behaviour, its frequency or the mean/total of a sensitive quantitative variable.

The present contribution aims at bringing together methodological and practical aspects of the indirect questioning approach. Specifically, the survey plan and the results of some real surveys about drug use and sexual behaviour will be discussed during the conference. It will be shown how the techniques employed in the surveys can enhance respondents’ cooperation and, according to the “more-is-better” principle, procure more reliable estimates than those stemming from traditional direct questioning (DQ) survey methods.

# Methods

The indirect questioning approach comprises various strategies for eliciting sensitive information, which may be mainly classified in three groups: the randomized response technique (RRT), the item count technique (ICT) and the nonrandomized response technique (NRRT). In terms of the volume of research conducted since Warner's [2] pioneering work, the RRT maintains a prominent, although controversial, position among the IQTs. Fundamentally, the RRT employs (at least in its original formulation) a physical randomization device (decks of cards, coloured numbered balls, dice, coins, spinners, random number generators, etc.) which determines whether respondents should answer the sensitive question or another, neutral one, or even provide a pre-specified response (e.g. “yes”) irrespective of their true status concerning the sensitive behaviour. The randomization device generates a probabilistic relation between the sensitive question and the answer released by the respondents, which is used to draw inferences on unknown population parameters of interest. The rationale of the RRT is that the respondents are less inhibited when the confidentiality of their responses is guaranteed. This goal is achieved because all responses are given according to the outcome of the randomization procedure, which is unknown to both the interviewer and the researcher and, hence, respondents' privacy is preserved. Similar protection is assured by the ICT originally proposed by Raghavarao and Federer [3]. Using this approach in surveys which pertain the study of a sensitive qualitative variable, two independent samples of respondents are selected according to a given sampling design. Without loss of generality, units belonging to the first sample are provided with a *long list* of items containing G+1 dichotomous questions, of which G are nonsensitive, while the remaining one refers to a sensitive attribute whose prevalence has to be estimated. Units in the second sample receive, instead, a *short list* containing only the G innocuous questions which are identical to those present in the long list. The sampled units are asked to report the total number of items that apply to them without revealing which item applies individually. Consequently, the privacy of the respondents provided with the long list is protected since their true sensitive status remains undisclosed unless they report that none or all of the G+1 items in the list apply to them. Finally, in the NRRT, no physical device is adopted, and neither are respondents asked to run a randomizing procedure. Instead, the respondents answer according to their true beliefs regarding the sensitive question and to one or more nonsensitive variables. For the discussion about some of the techniques ascribable to this approach see, e.g., Tian and Tang [4].

In the course of time, many variants of the original IQTs have been proposed in the literature with the aim of increasing the perceived level of privacy protection, improving the efficiency of estimation process and taking into account more complex survey situations. For instance, Rueda et al. [5] discussed estimation of sensitive parameters when data are collected using the RRT in multiple frames surveys. Focusing on the ICT, a variant of it, conceived for quantitative sensitive variables, has been introduced in Chaudhuri and Christofides [1] and termed the item sum technique (IST) in Trappmann et. al [6]. The IST works in a similar way to the ICT. Units belonging to the two samples are presented with the two lists of items which refer to quantitative variables, possibly measured on the same scale as that of the sensitive variable. Respondents are then asked to report the total score of their answers to all of the questions in their list, without revealing the individual score for each question. Like the ICT, the mean difference of the answers between the long list sample and the short list sample is used as an unbiased estimator of the population mean of the sensitive variable. Recent methodological advances of the IST have been provided by Rueda et al. [7] and Perri et al. [8], and mainly concern: (i) the use of auxiliary information at the design and estimation stages; (ii) the implementation of the technique when two or more sensitive variables are investigated and the estimation of their population means is required; (iii) the optimal allocation of the respondents in the short and long list samples.

Interested readers may found many other recent advances about IQTs in Chaudhuri et al. [9].

# Results

The IQTs have been used in many fields of applied research, particularly in sociological, economic, demographic, ecological and medical studies. Investigated topics include, for instance, the use of illegal drugs, athletic and cognitive performance-enhancing substances, the estimation of the prevalence of fraud in the area of disability benefits, tax evasion, plagiarism, assessment of sensitive health-risk behaviours in HIV/AIDS positive individuals, farmers' transgressionary behaviours and prevalence estimation of animal diseases such as sheep scab and swine fever, illegal bushmeat hunting, illegal fishing and unauthorized natural resources use. In the following, we mention some recent studies where different IQTs have been applied vis-à-vis DQ survey modes.

Perri et al. [10] employed the *crossed model* proposed by Lee at al. [11] in a pilot survey (868 women) to investigate the prevalence of two sensitive characteristics in Calabria (Italy), say induced abortion among foreign women residing in the region and irregular immigrant status. The effectiveness of the crossed model has also been investigated in another study (289 individuals) to simultaneously estimate the prevalence of individuals who smoke cannabis and support cannabis legalization. Perri et al. [12], on the basis of a stratified sample of 2398 university students, have estimated the prevalence and frequency of cannabis use and problematic sexual behaviours using the RRT and the IST. Hsieh and Perri [13] tested the Christofides [14] randomized response procedure in a face-to-face survey to estimate, under a Bayesian framework, non-heterosexual identity among Taiwanese individuals aged 20 years or older.

In general, all the surveys conducted have shown that the IQTs can really improve the estimates obtained when DQ survey techniques are used to elicit sensitive information. Detailed results are intentionally skipped for space saving purpose but they will be shown and discussed during the conference, possibly together with other experiences about the collection of sensitive data by means of NRRT procedures.

# References

1. A. Chaudhuri and T.C. Christofides, Indirect Questioning in Sample Surveys, Springer-Verlag, Berlin Heidelberg, 2013.
2. S.L. Warner, Randomized response: A survey technique for eliminating evasive answer bias, Journal of the American Statistical Association 60 (1965), 63-69.
3. D. Raghavarao and W.F. Federer, Block total response as an alternative to the randomized response method in survey, Journal of the Royal Statistical Society-B 41 (1979), 40-45.
4. G.-L. Tian and M.-L. Tang, Incomplete Categorical Data Design: Non-Randomized Response Techniques for Sensitive Questions in Surveys, Chapman & Hall/CRC, Boca Raton, FL, 2014.
5. M. Rueda, B. Cobo and P.F. Perri, Randomized response estimation in multiple frame surveys, International Journal of Computer Mathematics (2018), DOI 10.1080/00207160.2018.1476856, in press.
6. M. Trappmann, I. Krumpal, A Kirchner and B. Jann, Item sum: a new technique for asking quantitative sensitive questions, Journal of Survey Statistics and Methodology 2 (2014), 58-77.
7. M. Rueda, P.F. Perri and B. Cobo, Advances in estimation by the item sum technique using auxiliary information in complex surveys, AStA-Advances in Statistical Analysis (2017), DOI 10.1007/s10182-017-0315-2, in press.
8. P.F. Perri, M. Rueda and B. Cobo, Multiple sensitive estimation and optimal sample size allocation in the item sum technique, Biometrical Journal60 (2018), 155-163.
9. A. Chaudhuri, T.C. Christofides and C.R Rao (Eds), Data Gathering, Analysis and Protection of Privacy through Randomized Response Techniques: Qualitative and Quantitative Human Traits. Handbook of Statistics 34, Elsevier, Amsterdam, 2016.
10. P.F. Perri, E. Pelle and M. Stranges, Estimating induced abortion and foreign irregular presence using the randomized response crossed model, Social Indicators Research 129 (2015) 129, 601-618.
11. C.-S Lee., S.A. Sedory and S. Singh, Estimating at least seven measures of qualitative variables from a single sample using randomized response technique, Statistics and Probability Letters 83 (2013), 399-409.
12. P.F. Perri, B. Cobo and M. Rueda, A mixed-mode sensitive research on cannabis use and sexual addiction: improving self-reporting by means of indirect questioning techniques. Quality & Quantity (2017), DOI: 10.1007/s11135-017-0537-0, in press.
13. S.-H. Hsieh and P.F. Perri, Estimating non-heterosexuality identity in Taiwan using Christofides’ randomized response method: Some evidence from a real study, Submitted (2018).
14. T.C. Christofides, A generalized randomized response technique, Metrika 57 (2003), 195-200.

1. Department of Economics, Statistics and Finance – University of Calabria, Italy [↑](#footnote-ref-1)