Check for updates

Efficient item count techniques with one or two lists

Andreas Quatember[1] 📵

Abstract
If the direct questioning on sensitive variables leads to non-ignorable item-nonresponse and untruthful answering, a considerably biased estimator might be the consequence. In such cases, indirect questioning designs, which protect the respondents' privacy by masking the sensitive information, could pay off in terms of accuracy through an increased willingness to cooperate. To achieve this goal, such a design has to be simple in its implementation for the users and easy to understand for the respondents. In this article, it is shown for one of the indirect questioning designs, the item count technique, how the usage of specific oftentimes available prior information can substantially improve the estimation accuracy and at the same time reduce the respondents' task. This can make the method a stronger and more serious competitor of the direct questioning on sensitive attributes, which is commonly used in empirical research.

1 Introduction

When questions on sensitive topics, such as cyber bullying, illegal work, sexual behavior, or compulsory vaccination, are directly asked in statistical surveys, the rates of item-nonresponse as well as of untruthful answering might increase far above the usual levels because such questions can be seen as invasion of privacy, or certain answers on these questions can be considered to be socially unacceptable (cf., for instance, Tourangeau and Yan [26]). Such a behavior of the respondents might lead to strongly biased estimators of parameters under study such as the population proportion of people bearing a certain attribute. Therefore, before having to apply the methods of weighting adjustment or data imputation (cf., for instance, Särndal and Lundström [22]), which try to compensate just for the nonresponse that has occurred but not for the untruthful answering, everything should be done to increase the respondents' willingness to cooperate to make the rates of these two sources of systematic non-sampling errors as small as possible.

✉ Andreas Quatember
   andreas.quatember@jku.at

[1]   Johannes Kepler University (JKU) Linz, Science Park 2, Altenberger Str. 66a, 4040 Linz, Austria

Indirect questioning (IQ) designs intend to ensure respondents' cooperation by protecting their privacy. To achieve this goal, these techniques "mask" the respondents' actual status with respect to a variable under study (for an overview of IQ designs see, for instance, Chaudhuri and Christofides [4]). One of these methods is the item count technique (ICT), also known as unmatched count technique or list experiment. Its original version was discussed in detail by Droitcour et al. [9]. Using the ICT, when it comes to the sensitive question, the questionnaire shows a list of different statements, the "items," describing the membership of different population subgroups. Respondents are asked to report only the number of items that apply to them and not which of them apply. Two independent samples are drawn from the population of interest. In the control sample, the item list consists only of a number of so-called non-key items. In the other sample, the treatment sample, the list does additionally include the "key item" with respect to the sensitive membership of a certain population subgroup under study. The non-key items should be perceived by the respondents as meaningful information in the context of the questionnaire (Chaudhuri and Christofides [3], p. 592). The only task of the control sample is to deliver the information on the non-key items that is needed in the estimation process.

Compared with other privacy protecting IQ designs such as the "randomized response techniques," the main advantage of the ICT is that the task of the respondents can easily be understood without the need for complex instructions so that it can be implemented very simply even in self-administered questionnaires. Moreover, the interviewees do never have to supply the answer on the sensitive question directly. Various experiments examined the effectiveness of the method (cf., for instance, Droitcour et al. [9], Tsuchiya et al. [27], Coutts and Jann [8], Comsa and Postelnicu [7], Kiewiet de Jonge and Nickerson [16], Wolter and Laier [28], Blair et al. [2], and in particular the meta-analysis by Ehler et al. [10]). Exciting application examples can be found, for instance, in Comsa and Postelnicu [7], Malesky et al. [17], Frye et al. [11], Gibson et al. [12], Rinken et al. [21], or Wolter et al. [29].

Clearly, in order to be recognized as a serious competitor to the common direct questioning approach in empirical research, an IQ technique has to be easy to understand and implement and as accurate as possible. Furthermore, it should be applicable for general probability sampling because in surveys, in which sensitive questions are asked, oftentimes complex sampling methods including stratification and clustering are used. In Sect. 2 of this article, the statistical properties of the basic versions of the ICT with one or two lists of non-key items are discussed. In Sect. 3, modifications of these basic versions are proposed, which make use of available relevant information about at least a part of the used non-key items. These modifications aim to increase the accuracy of the survey results and at the same time reduce the respondents' burden in the questionnaire. The purpose of the calculations in Sect. 4 is to get a numerical impression of the possible positive effects of the application of the proposed ICT versions on the estimation accuracy. The article is concluded by a summary and an outlook to further research questions.

2 The item count technique

In a generalization of the original version of the ICT, two independent without-replacement probability samples s_1 and s_2 of sizes n_1 and n_2, respectively, are drawn from the study population U of size N by probability sampling methods S_1 and S_2 with first-order inclusion probabilities π_k and ρ_k, respectively, and second-order inclusion probabilities π_{kl} and ρ_{kl}.

Table 1 The original ICT

	Sample s_1:	Sample s_2:
Item list:	G non-key items (x) + key item (y)	G non-key items (x)
Response:	$z_k = x_k + y_k$	x_k

In the control sample s_2, the item list consists only of G non-key items. An example with G = 5 non-key items is:

- *I am an only child.*
- *I use an electric toothbrush.*
- *I have had a reported traffic accident last year.*
- *I have been hospitalized last year.*
- *I have been abroad in the last year.*

The answer x to be reported by a respondent k from the control sample s_2 be x_k, the number of the G non-key items that apply ($x_k = 0, 1,…,G$). In the treatment sample s_1, the short item list consisting of the G non-key items is complemented by the key item under study, which describes the sensitive membership of a certain population subgroup $U_A \subset U$. An example is:

- *I have been engaged in undeclared work in the last year.*

Let variable y indicate this membership of respondent k from the treatment sample s_1:

$$y_k = \begin{cases} 1 \ if \ k \ \in \ U_A \\ 0 \ otherwise. \end{cases}$$

The answer z to be actually reported by such a respondent be $z_k = x_k + y_k$, the number of all G + 1 items that apply from the long item list ($z_k = 0, 1,…, G, G + 1$) (Table 1). Let the proportion p of interest be given by

$$p = \frac{1}{N} \cdot \sum_U y_k \tag{1}$$

($\sum_U$ is an abbreviated notation for the sum over all units k ∈ U). Parameter p from Eq. (1) can be expressed by the difference of the population means μ_z and μ_x of variables z and x, respectively:

$$p = \frac{1}{N} \cdot \sum_U z_k - \frac{1}{N} \cdot \sum_U x_k = \mu_z - \mu_x \tag{2}$$

Consequently, the difference

$$\hat{p} = \frac{1}{N} \cdot \sum_{s_1} \frac{z_k}{\pi_k} - \frac{1}{N} \cdot \sum_{s_2} \frac{x_k}{\rho_k} = \overline{z}_{HT} - \overline{x}_{HT} \tag{3}$$

of the Horvitz–Thompson-based estimators $\overline{z}_{HT}$ and $\overline{x}_{HT}$ of the two population means μ_z and μ_x calculated from the probability samples s_1 and s_2, provides an unbiased moment estimator of p under the sampling designs S_1 and S_2 (cf. Särndal et al. [23], Sect. 2.8). The estimate $\hat{p}$ could be outside [0; 1]. For such cases, the maximum-likelihood estimator using the EM algorithm is a possible solution (see, for instance, Tian et al. [25]). If non-ignorable nonresponse still occurs, for the application of weighting adjustment techniques, see, for instance, Barabesi et al. [1].

The theoretical variance $V(\hat{p})$ of the estimator $\hat{p}$ from Eq. (3) is given by the sum of the usual theoretical variances $V(\overline{z}_{HT})$ and $V(\overline{x}_{HT})$ of $\overline{z}_{HT}$ and $\overline{x}_{HT}$, respectively:

$$V(\hat{p}) = V(\overline{z}_{HT}) + V(\overline{x}_{HT})$$
$$= \frac{1}{N^2} \cdot \left[\sum\sum_U (\pi_{kl} - \pi_k \cdot \pi_l) \cdot \frac{z_k}{\pi_k} \cdot \frac{z_l}{\pi_l} + \sum\sum_U (\rho_{kl} - \rho_k \cdot \rho_l) \cdot \frac{x_k}{\rho_k} \cdot \frac{x_l}{\rho_l} \right]$$
(4)

($\sum\sum_U$ is an abbreviated notation for the double sum over all units k,l $\in$ U). For the effect of the selection of the non-key items on $V(\hat{p})$, see, for instance, Glynn [13]. For the variance-optimal allocation of the total sample size n on the two samples, see, for instance, Tian et al. [25]. Perri et al. [19] discuss the idea of optimal allocation extensively for the Item Sum Technique.

The variance from Eq. (4) can be unbiasedly estimated by

$$\hat{V}(\hat{p}) = \hat{V}(\overline{z}_{HT}) + \hat{V}(\overline{x}_{HT})$$
$$= \frac{1}{N^2} \cdot \left[\sum\sum_{s_1} \frac{\pi_{kl} - \pi_k \cdot \pi_l}{\pi_{kl}} \cdot \frac{z_k}{\pi_k} \cdot \frac{z_l}{\pi_l} + \sum\sum_{s_2} \frac{\rho_{kl} - \rho_k \cdot \rho_l}{\rho_{kl}} \cdot \frac{x_k}{\rho_k} \cdot \frac{x_l}{\rho_l} \right]$$
(5)

For simple random sampling with replacement (SIR) (or also approximately for simple random sampling without replacement from large populations) in both samples, for instance, Eq. (3) results in

$$\hat{p}_{SIR} = \overline{z} - \overline{x},$$
(6)

the difference of the simple sample means of z and x in the two SIR samples s_1 and s_2, respectively. For this sampling design, Eq. (4) results in

$$V(\hat{p}_{SIR}) = \frac{\sigma_z^2}{n_1} + \frac{\sigma_x^2}{n_2}$$
(7)

with σ_z^2 and σ_x^2, the population variances of z and x. Eventually, Eq. (5) yields

$$\hat{V}(\hat{p}_{SIR}) = \frac{s_z^2}{n_1} + \frac{s_x^2}{n_2}$$
(8)

with s_z^2 and s_x^2, respectively, the sample variances of z and x in s_1 and s_2.

Besides the previously mentioned advantages of this procedure, also two weaknesses have been discussed in the relevant literature. One is the waste of estimation accuracy because the sensitive information under study is observed only in the treatment sample s_1, whereas the control sample s_2 only serves as a reference for the calculation of the estimate $\overline{x}_{HT}$ needed in Eq. (3). The other one is that in s_1, the process answers $z_k = G + 1$ and $z_k = 0$, respectively, do reveal a respondent's true status on the sensitive membership of the group U_A (and at the same time also of all non-key items). The probabilities of the occurrence of these "ceiling" or "floor effects," respectively, can be reduced by a proper choice of the non-key items (see, for instance, Glynn [13], p. 163). The floor effect is less problematic than the ceiling effect unless also the non-membership of U_A is sensitive. In s_2, the same applies to the response answers $x_k = G$ and $x_k = 0$ with respect to the non-key items.

As a consequence of these weaknesses, several modifications of this original ICT have been proposed. The double-list (DL) version by Droitcour et al. [9], for instance, addresses the waste of efficiency by adding to the questionnaires in both samples of the original ICT a second item list, where the non-key items of the first list are replaced by other non-key

Table 2 The DL version of the ICT

	Sample s_1:	Sample s_2:
First item list:	G non-key items (x) + key item (y)	G non-key items (x)
First Response:	$z_k = x_k + y_k$	x_k
Second item list:	G other non-key items (u)	G other non-key items (u) + key item (y)
Second Response:	u_k	$w_k = u_k + y_k$

items. Without loss of generality, let us assume that the number of non-key items equals G in both lists. But, the treatment sample s_1 of the original ICT with respect to the first list serves now at the same time as the control sample with respect to the second list and vice versa. Therefore, two answers, z and u, have to be given by a respondent k from sample s_1. These are the number $z_k = x_k + y_k$ of applicable items from the long first item list including the key item (as in the original ICT) and the number u_k of applicable non-key items from the short second list without the key item ($u_k = 0, 1,…, G$) (Table 2). The answers x and w to be given by a respondent k from sample s_2 are the number x_k of applicable non-key items from the short first list (as in the original ICT) and the number $w_k = u_k + y_k$ of applicable items from the long second list ($w_k = 0, 1,…, G, G + 1$).

By this supplement to the questionnaires, in both samples information on the sensitive variable is observed. This increases the estimation accuracy compared with the original ICT by the price of an only insignificant increase in the respondents' burden by the usage of a second item list that should hardly negatively affect their willingness to cooperate.

With two different lists of non-key items applied to two samples, from Eq. (3) two separate estimates $\hat{p}_1$ and $\hat{p}_2$, respectively, can be calculated, of which their mean value

$$\overline{\hat{p}} = \frac{\hat{p}_1 + \hat{p}_2}{2} = \frac{1}{2} \cdot \left(\frac{1}{N} \cdot \sum_{s_1} \frac{z_k}{\pi_k} - \frac{1}{N} \cdot \sum_{s_2} \frac{x_k}{\rho_k} + \frac{1}{N} \cdot \sum_{s_2} \frac{w_k}{\rho_k} - \frac{1}{N} \cdot \sum_{s_1} \frac{u_k}{\pi_k} \right)$$
$$= \frac{\overline{z}_{HT} - \overline{x}_{HT} + \overline{w}_{HT} - \overline{u}_{HT}}{2} \tag{9}$$

is taken as the procedure's unbiased estimate. In Eq. (9), $\overline{z}_{HT}$ and $\overline{u}_{HT}$ calculated from the probability sample s_1, and $\overline{x}_{HT}$ and $\overline{w}_{HT}$ calculated from the probability sample s_2, respectively, are the Horvitz-Thompson-based estimators of the population means μ_z, μ_u, μ_x, and μ_w. The theoretical variance of $\overline{\hat{p}}$ is given by

$$V(\overline{\hat{p}}) = \frac{1}{4} \cdot \left[V(\hat{p}_1) + V(\hat{p}_2) + 2 \cdot C(\hat{p}_1, \hat{p}_2) \right] \tag{10}$$

Equation (10) includes the two variances $V(\hat{p}_1)$ and $V(\hat{p}_2)$, which are calculated applying Eq. (4), and the covariance term $C(\hat{p}_1, \hat{p}_2)$, addressing the levels of dependence of z and u in s_1, and x and w in s_2, respectively (for the details see Appendix 1). The formula for the estimator of this variance includes the variance estimators that can be generated straightforward from Eq. (5) and an estimator of the covariance term from the sample data. The relevant formulas under the SIR sampling design with the estimator

$$\overline{\hat{p}}_{SIR} = \frac{\hat{p}_{1,SIR} + \hat{p}_{2,SIR}}{2} \tag{11}$$

can be derived accordingly.

Petróczi et al. [20] and Groenitz [14] proposed the "single sample count technique." For this ICT approach with only one list, the joint population distribution of the non-key items is assumed to be known so that a control sample is no longer needed. Its theory is developed for the SIR design under practically limiting assumptions about the independence of the individual items. Like the DL version, this technique addresses the efficiency problem of the original ICT. The modifications of the original ICT that will be proposed in Sect. 3 are based on these contributions.

Other modifications of the original ICT concern its ceiling/floor effect. Such methods were presented by Chaudhuri and Christofides [3], Christofides [5], Ibrahim [15], Shaw [24], Christofides and Manoli [6], or Manoli [18].

3 The proposed modifications of the original item count technique

In this section, a modification of the original ICT is proposed that increases the accuracy of the estimation of the proportion p and at the same time reduces the complexity of the questioning design. For this modification, we presume that the population mean value of the number of applicable non-key items among at least a part of these items is available, for instance, from administrative or register data. Non-key items for which this information is available and which do not appear to be completely meaningless, such as asking for the last digit of the phone number, should not be too difficult to find. For this purpose, for example, socio-demographic items with known distribution in the target population such as marital status (e.g. the statement "*I am unmarried.*") or education (e.g. "*I have a degree of a university of applied sciences.*") might be used. Other examples of such items could be age, gender, place of residence, migration background, nationality, ethnicity, religion, household size, employment, income or working hours. In the following, the effect of the usage of such information on the expressions of the estimator, its theoretical variance, and the variance estimator, respectively, is presented for general probability sampling.

Let the population mean value $\mu_{x^{(F_1)}}$ of the number $x^{(F_1)}$ of applicable items among F_1 of the G non-key items be given ($0 \leq F_1 \leq G$). Hence, the expression of the parameter p from Eq. (2) can be re-written by

$$p = \mu_z - \left(\mu_{x^{[E_1]}} + \mu_{x^{(F_1)}}\right) \tag{12}$$

with $\mu_{x^{[E_1]}}$, the unknown population mean value of the number $x^{[E_1]}$ of applicable items among the $E_1 = G - F_1$ of the G non-key items that are not contained in $\mu_{x^{(F_1)}}$.

For $F_1 = 0$ ($E_1 = G$), the item list corresponds to that of the original ICT. But for $0 < F_1 < G$, the item list of the control sample s_2 consists only of the E_1 non-key items that are not included in $\mu_{x^{(F_1)}}$, which reduces the respondents' task in sample s_2 compared with the original ICT. The answer to be reported by a respondent k from sample s_2 is $x_k^{[E_1]}$, the number of the E_1 non-key items that apply ($x_k^{[E_1]} = 0, 1,\ldots, E_1$). In the treatment sample s_1, the answer to be reported by a respondent is the same as in the original ICT (Table 3).

Clearly, the parameter from Eq. (12) is unbiasedly estimated by

$$\hat{p}^{(F_1)} = \overline{z}_{HT} - \left(\overline{x}_{HT}^{[E_1]} + \mu_{x^{(F_1)}}\right) \tag{13}$$

with $\overline{z}_{HT}$ from Eq. (3) and the Horvitz–Thompson-based estimator

$$\overline{x}_{HT}^{[E_1]} = \frac{1}{N} \cdot \sum_{s_2} \frac{x_k^{[E_1]}}{\rho_k}$$

Table 3 The proposed modified version of the original ICT

	Sample s_1:	Sample s_2:
Item list:	G non-key items (x) + key item (y)	E_1 non-key items ($x^{[E_1]}$)
Response:	$z_k = x_k + y_k$	$x_k^{[E_1]}$

of the population mean $\mu_{x^{[E_1]}}$ from sample s_2. For $F_1 = 0$, $\hat{p}^{(0)} = \hat{p}$ from Eq. (3) applies.

The theoretical variance $V(\hat{p}^{(F_1)})$ of the estimator $\hat{p}^{(F_1)}$ from Eq. (13) is given by the sum of the two theoretical variances $V(\bar{z}_{HT})$ and $V(\bar{x}_{HT}^{[E_1]})$ of $\bar{z}_{HT}$ and $\bar{x}_{HT}^{[E_1]}$, respectively,

$$V(\hat{p}^{(F_1)}) = V(\bar{z}_{HT}) + V\left(\bar{x}_{HT}^{[E_1]}\right)$$
$$= \frac{1}{N^2} \cdot \left[\sum\sum_U (\pi_{kl} - \pi_k \cdot \pi_l) \cdot \frac{z_k}{\pi_k} \cdot \frac{z_l}{\pi_l} + \sum\sum_U (\rho_{kl} - \rho_k \cdot \rho_l) \cdot \frac{x_k^{[E_1]}}{\rho_k} \cdot \frac{x_l^{[E_1]}}{\rho_l} \right] \quad (14)$$

The accuracy of the estimator $\hat{p}^{(F_1)}$ increases with $F_1 \to G$.

This variance is unbiasedly estimated by

$$\hat{V}(\hat{p}^{(F_1)}) = \hat{V}(\bar{z}_{HT}) + \hat{V}\left(\bar{x}_{HT}^{[E_1]}\right)$$
$$= \frac{1}{N^2} \cdot \left[\sum\sum_{s_1} \frac{\pi_{kl} - \pi_k \cdot \pi_l}{\pi_{kl}} \cdot \frac{z_k}{\pi_k} \cdot \frac{z_l}{\pi_l} + \sum\sum_{s_2} \frac{\rho_{kl} - \rho_k \cdot \rho_l}{\rho_{kl}} \cdot \frac{x_k^{[E_1]}}{\rho_k} \cdot \frac{x_l^{[E_1]}}{\rho_l} \right] \quad (15)$$

Under the SIR design in both samples s_1 and s_2, for instance, Eq. (13) results in

$$\hat{p}_{SIR}^{(F_1)} = \bar{z} - (\bar{x}^{[E_1]} + \mu_{x^{(F_1)}}) \quad (16)$$

with the sample means $\bar{z}$ and $\bar{x}^{[E_1]}$ of z and $x^{[E_1]}$ in s_1 and s_2, respectively. For this sampling design, Eq. (14) results in

$$V(\hat{p}_{SIR}^{(F_1)}) = \frac{\sigma_z^2}{n_1} + \frac{\sigma_{x^{[E_1]}}^2}{n_2} \quad (17)$$

with σ_z^2 and $\sigma_{x^{[E_1]}}^2$, the population variances of z and $x^{[E_1]}$. Eventually, Eq. (15) yields

$$\hat{V}(\hat{p}_{SIR}^{(F_1)}) = \frac{s_z^2}{n_1} + \frac{s_{x^{[E_1]}}^2}{n_2} \quad (18)$$

with s_z^2 and $s_{x^{[E_1]}}^2$, respectively, the sample variances of z and $x^{[E_1]}$ in the SIR samples s_1 and s_2.

Following the idea of Petróczi et al. [20], in the special case that $F_1 = G$, a control sample is no longer needed because the mean number μ_x of x from Eq. (2) is known ($\mu_{x^{(G)}} \equiv \mu_x$, $x^{(G)} \equiv x$). This means that the total sample number $n = n_1 + n_2$ can be allocated to the treatment sample alone. With only one long list in the whole sample of size n, Eq. (13) reduces to

$$\hat{p}^{(G)} = \bar{z}_{HT} - \mu_x \quad (19)$$

and Eqs. (14) and (15), respectively, to

$$V(\hat{p}^{(G)}) = V(\bar{z}_{HT}) = \frac{1}{N^2} \cdot \left[\sum\sum_U (\pi_{kl} - \pi_k \cdot \pi_l) \cdot \frac{z_k}{\pi_k} \cdot \frac{z_l}{\pi_l} \right] \tag{20}$$

and

$$\hat{V}(\hat{p}^{(G)}) = \hat{V}(\bar{z}_{HT}) = \frac{1}{N^2} \cdot \left[\sum\sum_{s_1} \frac{\pi_{kl} - \pi_k \cdot \pi_l}{\pi_{kl}} \cdot \frac{z_k}{\pi_k} \cdot \frac{z_l}{\pi_l} \right] \tag{21}$$

For the SIR method, these equations yield

$$\hat{p}_{SIR}^{(G)} = \bar{z} - \mu_x, \tag{22}$$

$$V(\hat{p}_{SIR}^{(G)}) = \frac{\sigma_z^2}{n} \tag{23}$$

and

$$\hat{V}(\hat{p}_{SIR}^{(G)}) = \frac{s_z^2}{n} \tag{24}$$

For a combination of this modified ICT, which uses relevant prior information on the non-key items, with the DL version of the ICT, which uses two different item lists, let the mean values $\mu_{x^{(F_1)}}$ of the number $x^{(F_1)}$ of applicable items of F_1 of the G non-key items from the first list ($0 \leq F_1 \leq G$) and $\mu_{u^{(F_2)}}$ of the number $u^{(F_2)}$ of applicable items of F_2 of the G non-key items from the second list ($0 \leq F_2 \leq G$) be known. In this case, the short first item list included in sample s_2 consists only of the $E_1 = G - F_1$ non-key items that are not included in the known mean value $\mu_{x^{(F_1)}}$, whereas the short second item list included in sample s_1 consists only of the $E_2 = G - F_2$ non-key items that are not included in the known mean value $\mu_{u^{(F_2)}}$. Hence, for $F_1 > 0$ and/or $F_2 > 0$, the use of such prior information does also reduce the respondents' burden of the original DL design. The numbers to be reported by a respondent k from sample s_1 are z_k, the number of applicable items from the long first item list including the key item, and $u_k^{[E_2]}$, the number of applicable non-key items from the short second item list ($u_k^{[E_2]} = 0, 1,\ldots, E_2$). A respondent k from sample s_2 has to report $x_k^{[E_1]}$, the number of applicable non-key items from the short first item list ($x_k^{[E_1]} = 0, 1,\ldots, E_1$), and $w_k = u_k + y_k$, the number of applicable items from the long second item list (Table 4). For $F_1 = F_2 = 0$, the item lists correspond to those of the original DL version.

Table 4 The modified DL version of the ICT		Sample s_1:	Sample s_2:
First item list:		G non-key items (x) + key item (y)	E_1 non-key items ($x^{[E_1]}$)
First Response:		$z_k = x_k + y_k$	$x_k^{[E_1]}$
Second item list:		E_2 other non-key items ($u^{[E_2]}$)	G other non-key items (u) + key item (y)
Second Response:		$u_k^{[E_2]}$	$w_k = u_k + y_k$

This questioning design changes the expression of the estimator (9) to

$$\overline{\hat{p}}^{(F_1/F_2)} = \frac{\hat{p}_1^{(F_1)} + \hat{p}_2^{(F_2)}}{2} \qquad (25)$$

This is the mean value of the two separate estimates $\hat{p}_1^{(F_1)}$ and $\hat{p}_2^{(F_2)}$ that can be retrieved from Eq. (13). Its theoretical variance is given by

$$V(\overline{\hat{p}}^{(F_1/F_2)}) = \frac{1}{4} \cdot [V(\hat{p}_1^{(F_1)}) + V(\hat{p}_2^{(F_2)}) + 2 \cdot C(\hat{p}_1^{(F_1)}, \hat{p}_2^{(F_2)})] \qquad (26)$$

with the two variances $V(\hat{p}_1^{(F_1)})$ and $V(\hat{p}_2^{(F_2)})$, respectively, according to Eq. (14) and a covariance term $C(\hat{p}_1^{(F_1)}, \hat{p}_2^{(F_2)})$ that addresses the levels of dependence of z and $u^{[E_2]}$ in s_1, and $x^{[E_1]}$ and w in s_2, respectively. The variance term (26) can be estimated applying Eq. (21) for the estimation of the variance terms and an estimator of the covariance term using the sample data. For the SIR design, Eq. (25) yields

$$\overline{\hat{p}}_{SIR}^{(F_1/F_2)} = \frac{\hat{p}_{1,SIR}^{(F_1)} + \hat{p}_{2,SIR}^{(F_2)}}{2} \qquad (27)$$

If one or both mean values μ_x and μ_u are known ($F_1 = G$ and/or $F_2 = G$), for the first, the second or both item lists, a control sample would theoretically no longer be needed. But, to apply one of the two or even both long lists to the entire sample of size n could be counter-productive because in such cases, the members of at least one sample would have to respond to two long lists with the same key item. This might have a negative impact on their perceived privacy protection and consequently on their cooperation willingness.

4 A Numerical comparison of the accuracy of different versions of the ICT

In this section, the different ICT versions described in Sects. 2 and 3 are numerically compared under the SIR design to provide an impression of the effect of the proposed techniques on the accuracy of the estimation. For all these methods, the respondents' task is of similar simplicity, which should ensure a similar effect on the cooperation willingness. For this purpose, from the "imaginary population" provided by Shaw [24], the six-dimensional population distribution of the $G = 5$ non-key items (there named "1" to "5") and the key item (named "A"), and their dependence structure is used (see Appendix 2).

On the one hand, for sample sizes n_1 and n_2, the estimation accuracy of the estimators $\hat{\theta}_{SIR}$ ($\hat{p}_{SIR}^{(F_1)}$ from Eq. (16) and $\overline{\hat{p}}_{SIR}^{(F_1,F_2)}$ from Eq. (27), $0 \le F_1, F_2 \le G$) of the different ICT methods is compared by the relative variance reduction (RVR) in % of $V(\hat{p}_{SIR})$ of the original ICT (Table 5):

$$RVR = \left(1 - \frac{V(\hat{\theta}_{SIR})}{V(\hat{p}_{SIR})}\right) \cdot 100 \qquad (28)$$

For $\hat{\theta}_{SIR} = \hat{p}_{SIR}^{(5)}$ from Eq. (22), for example, for which not only the population mean μ_x is known, but also the variable z is observed in the whole sample of size $n = 1{,}000$, with the

Table 5 RVR in % compared with $V(\hat{p}_{SIR})$ of the original ICT

$F_1 \backslash F_2$	$\hat{p}_{SIR}^{(F_1)}$	$\overline{\hat{p}}_{SIR}^{(F_1/F_2)}$					
		0	1	2	3	4	5
0	0	49.2					
1	11.2	52.1	55.1				
2	20.5	53.1	56.0	58.4			
3	26.7	54.5	57.4	59.8	62.0		
4	35.4	57.6	60.5	62.9	65.1	67.2	
5	72.7	59.4	62.3	64.7	66.9	69.0	71.0

variances $V(\hat{p}_{SIR}^{(5)})$ and $V(\hat{p}_{SIR})$ from Eqs. (20) and (7), the RVR in % is given by

$$RVR = \left(1 - \frac{\frac{\sigma_z^2}{n}}{\frac{\sigma_z^2}{n_1} + \frac{\sigma_x^2}{n_2}}\right) \cdot 100$$

With the parameters from Shaw's population (see Appendix 2), this results in an RVR of 72.7% compared to the original ICT approach (Table 5).

On the other hand, the estimation accuracy of the estimators $\hat{\theta}_{SIR}$ is compared with that of the "direct" estimator

$$\hat{p}_{SIR}^{D} = \frac{1}{n} \cdot \sum_{s} r_k, \tag{29}$$

the proportion of the "yes-"responses $r = 1$ in the SIR sample of size $n = n_1 + n_2$ with

$$r_k = \begin{cases} 1 \ if \ respondent \ k \ answers \ ''yes'' \\ 0 \ if \ respondent \ k \ answers \ ''no,'' \end{cases}$$

assuming that with probability q instead of the true status $y = 1$ the response $r = 0$ is reported, when the sensitive question is asked directly, and that all other non-sampling error components are equally negligible for the considered questioning designs. For this purpose, the threshold value q_0 of the probability q is calculated, the exceeding of which yields a mean square error $MSE(\hat{p}_{SIR}^{D})$ larger than the variance $V(\hat{\theta}_{SIR})$ of the unbiased estimator $\hat{\theta}_{SIR}$ (Table 6):

$$q_0 = \frac{\sqrt{V(\hat{\theta}_{SIR}) - \frac{p(1-p)}{n}}}{p} \tag{30}$$

(for the theoretical development, see Appendix 3). For $q > q_0$, the bias of $\hat{p}_{SIR}^{D}$ will yield $MSE(\hat{p}_{SIR}^{D}) > V(\hat{\theta}_{SIR})$, and the increased privacy protection of the specific ICT design will pay off in terms of accuracy. For $\hat{\theta}_{SIR} = \hat{p}_{SIR}^{(5)}$ from Eq. (22), for example, Eq. (30) results in

$$q_0 = \frac{\sqrt{\frac{\sigma_z^2}{n} - \frac{p(1-p)}{n}}}{p}$$

Table 6 q_0 for which $MSE(\hat{p}_{SIR}^D) \leq V(\hat{\theta}_{SIR})$ applies

$F_1 \backslash F_2$	$\hat{p}_{SIR}^{(F_1)}$	$\overline{p}_{SIR}^{(F_1/F_2)}$					
		0	1	2	3	4	5
0	0.144	0.100					
1	0.135	0.097	0.093				
2	0.128	0.096	0.092	0.089			
3	0,122	0.094	0.091	0.088	0.085		
4	0.114	0.090	0.087	0.084	0.081	0.078	
5	0.070	0.088	0.084	0.081	0.078	0.075	0.072

With the parameters from Shaw's population (see Appendix 2), this results in $q_0 = 0.070$ (see Table 6). Under the given assumptions, for a probability q larger than only 7.0%, $V(\hat{p}_{SIR}^{(5)}) < MSE(\hat{p}_{SIR}^D)$ applies and the modification of the original ICT with G = 5 will provide more accurate results than the direct questioning approach.

In the given data set, the proportion p under study is equal to 0.479. The comparison is done with $n_1 = n_2 = 500$. The uniform allocation of n to the two samples is a reasonable choice when DL versions of the ICT are included in the investigations. The results presented in Tables 5 and 6 provide a numerical impression of the possible effects of the knowledge of the mean values of at least a part of the G non-key-items in techniques with one or two item lists on the estimation accuracy. For the one list versions of the ICT, $n_1 > n_2$ would be a better choice than $n_1 = n_2$. Hence, for $\hat{p}_{SIR}^{(F_1)}$ ($0 \leq F_1 \leq 5$) the results in Table 5 can be interpreted as lower limits of the achievable RVR in % and the results in Table 6 as upper limits of the probability q_0.

Clearly, in a one-list design (column $\hat{p}_{SIR}^{(F_1)}$ in the Tables 5 and 6), the higher the number F_1 of the G non-key items is, for which $\mu_{x^{(F_1)}}$ is known, the higher is the RVR of $\hat{p}_{SIR}^{(F_1)}$ in comparison to $\hat{p}_{SIR}$ of the original ICT. For the given data (see Appendix 2), the DL approach ($\overline{p}_{SIR}^{(F_1/F_2)}$) with $F_1 = F_2 = 0$ ($\overline{p}_{SIR}$) roughly halves the variance of the original ICT ($\hat{p}_{SIR}$) by the use of two lists instead of only one. Moreover, Table 5 shows how also the DL version can gain additional accuracy through knowledge of the mean values $\mu_{x^{(F_1)}}$ and $\mu_{u^{(F_2)}}$, respectively ($0 \leq F_2 \leq F_1 \leq G$). This shows, how the performances of the estimators benefit from both ideas, use of prior knowledge and of two lists. Each additional part of prior information increases the accuracy of the estimator. In the applied setting, only the estimator $\overline{p}_{SIR}^{(5/5)}$ of the DL version with $F_1 = F_2 = 5$ provides roughly the same result as $\hat{p}^{(5)}$, in which no control sample is needed anymore.

When it comes to the comparison with the direct questioning on the sensitive item by q_0 from Eq. (30), Table 6 shows that the original ICT would pay off in terms of accuracy if more than only 14.4% of the members of U_A would lie to the sensitive question when they were asked directly. As a consequence of the achievable variance reductions of the proposed modified versions of the ICT presented in Table 5, this q_0 can accordingly be reduced. For the best case with respect to simplicity and accuracy of the questioning design (Table 5), the estimator $\hat{p}_{SIR}^{(5)}$ would already be more accurate than the direct estimator (29) (both with n = 1,000) if the absolute bias of $\hat{p}_{SIR}^D$ from Eq. (29) is larger than only $0.070 \cdot 0.479 = 0.033$.

5 Summary and outlook

The original ICT is an easy-to-understand and simple-to-implement IQ design to mask sensitive information with the aim to increase respondents' cooperation willingness. The take-away message of this article is that the usage of certain prior information about at least some of the involved non-key items can substantially decrease the estimators' additional variance. In addition to contextual non-key items for which such information might be available, for example, socio-demographic items with known distribution in the target population oftentimes could be used for this purpose. In this way, the method has the potential to become an even stronger and more serious competitor of the direct questioning about sensitive characteristics.

For $1 \leq F_1 \leq G-1$, this prior information could additionally be used for regression-type estimators. For this purpose, in the control sample two separate lists with E_1 and F_1 non-key items would have to be used. For large correlations of $x^{[E_1]}$ and $x^{(F_1)}$, such estimators could reduce the second component of the variance $V(\hat{p}^{(F_1)})$ from (14) while the unchanged first component remains responsible for the larger part of the sum. Based on the proposed modifications of the original ICT, it may be of interest to investigate the pros and cons of such estimators in some detail in possible future research.

Moreover, a further combination of the proposed modifications of the ICT with methods that account for the floor/ceiling effect could also address these weaknesses of the original ICT in protecting also the privacy of the few respondents to whom these effects would apply. However, such methods, of course, would have to maintain the simplicity of the original ICT to avoid an increase in the respondents' task. If this cannot be ensured, it would be preferable to reduce the probability of these two effects by selecting the non-key items appropriately.

Data availability The data presented in this study are available on request from the corresponding author.

Declarations

Conflict of interest There are no conflicts of interest directly or indirectly related to the work submitted for publication.

Appendix 1

Expression (10) is developed further in the following way:

$$V(\overline{\hat{p}}) = \frac{1}{4} \cdot \left[V(\hat{p}_1) + V(\hat{p}_2) + 2 \cdot C(\hat{p}_1, \hat{p}_2) \right]$$

$$= \frac{1}{4} \cdot [V(\overline{z}_{HT}) + V(\overline{x}_{HT}) + V(\overline{w}_{HT}) + V(\overline{u}_{HT}) + 2 \cdot C(\overline{z}_{HT} - \overline{x}_{HT}, \overline{w}_{HT} - \overline{u}_{HT})]$$

According to

$C(a - b, c - d) = C(a, c) - C(b, c) - C(a, d) + C(b, d)$, the covariance term on the right-hand side of the equation is decomposed into:

$$C(\overline{z}_{HT} - \overline{x}_{HT}, \overline{w}_{HT} - \overline{u}_{HT}) = -[C(\overline{z}_{HT}, \overline{u}_{HT}) + C(\overline{x}_{HT}, \overline{w}_{HT})]$$

Appendix 2

For the purpose of providing a numerical impression of the effect of the different modifications of the original ICT proposed in Sect. 3 on the accuracy of the estimates, not the small "imaginary population" from Shaw [24] itself is used but the multi-dimensional population distribution of the five non-key items and the key item from this data set. From this distribution, SIR-samples of sizes $n_1 = n_2 = 500$ are drawn. Regarding the knowledge of the mean values $\mu_{x^{(1)}}$, $\mu_{x^{(2)}}$, ... , $\mu_{x^{(5)}}$ of the numbers $x^{(1)}$, $x^{(2)}$,..., $x^{(5)}$ of applicable non-key items, the variables "1" to "5" from the data set are used in reverse order (the knowledge of $\mu_{x^{(1)}}$ means that the population mean value of variable "5" is known, whereas the mean value $\mu_{x^{[4]}}$ of the number of applicable non-key items "1" to "4" is not known; the knowledge of $\mu_{x^{(2)}}$ means that the mean value of the sum of variables "4" and "5" is known, whereas the mean value $\mu_{x^{[3]}}$ of the number of applicable non-key items "1" to "3" is not known; and so forth). To allow also calculations for the different DL versions of the proposed modifications, the variable values of the different five non-key items included in Shaw's population are additionally randomly ordered for the generation of a second item list. Furthermore, without loss of generality, $F_1 \geq F_2$ was assumed.

The prevalence rate of the sensitive attribute in the whole population is given by p = 0.479. The probability distribution of the number x = {0, 1, 2, 3, 4, 5} of non-key-items that apply is given by {0.017, 0.154, 0.282, 0.359, 0.162, 0.026}. The conditional prevalence distribution of y, given x = {0, 1, 2, 3, 4, 5}, yields {1, 0.$\dot{4}$, 0.424, 0.524, 0.474, 0.$\dot{3}$}. The probability distribution of the number u = {0, 1, 2, 3, 4, 5} of non-key-items that apply in the second list (if there is one) is given by {0.026, 0.154, 0.299, 0.299, 0.188, 0.034}. The conditional prevalence distribution of y, given u = {0, 1, 2, 3, 4, 5}, yields {0.$\dot{3}$, 0.5, 0.486, 0.514, 0.364, 0.75}.

The parameters needed for the calculations are $\sigma_z^2 = 1.365$, $\sigma_{x^{(1)}}^2 = 0.250$, $\sigma_{x^{(2)}}^2 = 0.467$, $\sigma_{x^{(3)}}^2 = 0.622$, $\sigma_{x^{(4)}}^2 = 0.854$, and $\sigma_x^2 = 1.134$; C(z,u) = 0.065, C(x,w) = 0.056, C(x$^{[4]}$,w) = 0.064, C(z,u$^{[4]}$) = 0.089, C(x$^{[3]}$,w) = − 0.003, C(z,u$^{[3]}$) = 0.082, C(x$^{[2]}$,w) = − 0.012, C(z,u$^{[2]}$) = 0.057, C(x$^{[1]}$,w) = 0.035, and C(z,u$^{[1]}$) = 0.026.

Appendix 3

The mean square error $MSE(\hat{p}_{SIR}^D)$ of $\hat{p}_{SIR}^D$ from Eq. (29) is given by

$$MSE(\hat{p}_{SIR}^D) = B^2(\hat{p}_{SIR}^D) + V(\hat{p}_{SIR}^D)$$

(cf., for instance, Särndal et al. [23], p. 40) with the squared bias

$$B^2(\hat{p}_{SIR}^D) = \left[E(\hat{p}_{SIR}^D) - p \right]^2$$

of $\hat{p}_{SIR}^D$ and the variance

$$V(\hat{p}_{SIR}^D) = \frac{p \cdot (1-p)}{n}$$

With the probability q of untruthful answering, given y = 1,

$$E(\hat{p}_{SIR}^D) = (1-q) \cdot p$$

applies. Hence, under the given assumptions

$$V(\hat{\theta}_{SIR}) < MSE(\hat{p}_{SIR}^D)$$

applies if

$$V(\hat{\theta}_{SIR}) < q^2 \cdot p^2 + \frac{p \cdot (1-p)}{n}$$

With respect to the threshold value q_0, this leads to the result from Eq. (30).

References

1. Barabesi, L., Diana, G., Perri, P.F.: Horvitz-Thompson estimation with randomized response and nonresponse. Model. Assist. Stat. Appl. **9**, 3–10 (2014). https://doi.org/10.3233/MAS-130274
2. Blair, G., Coppock, A., Moor, M.: When to worry about sensitivity bias: a social reference theory and evidence from 30 years of list experiments. Am. Polit. Sci. Rev. **114**, 1297–1315 (2020). https://doi.org/10.1017/S0003055420000374
3. Chaudhuri, A., Christofides, T.C.: Item Count Technique in estimating the proportion of people with a sensitive feature. J. Stat. Plan. Inference **137**, 589–593 (2007). https://doi.org/10.1016/j.jspi.2006.01.004
4. Chaudhuri, A., Christofides, T.C.: Indirect Questioning in Sample Surveys. Springer, Heidelberg (2013)
5. Christofides, T.C.: A new version of the item count technique. Model Assist. Stat. Appl. **10**, 289–297 (2015). https://doi.org/10.3233/MAS-150333
6. Christofides, T.C., Manoli, E.: Item count technique with no floor and ceiling effects. Commun. Stat.-Theory Methods **49**(6), 1330–1356 (2020). https://doi.org/10.1080/03610926.2018.1563165
7. Comsa, M., Postelnicu, C.: Measuring social desirability effects on self-reported turnout using the item count technique. Int. J. Public Opin. Res. **25**(2), 153–172 (2013). https://doi.org/10.1093/IJPOR/EDS019
8. Coutts, E., Jann, B.: Sensitive questions in online surveys: experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). Sociol. Methods Res. **40**(1), 169–193 (2011). https://doi.org/10.1177/0049124110390768
9. Droitcour, J., Caspar, R.A., Hubbard, M.L., Ezzati, T.M.: Chapter 11, The item count technique as a method of indirect questioning: a review of its development and a case study application. In: Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N., Sudman, S. (eds.) Measurement Errors in Surveys. Wiley, Hoboken (1991)
10. Ehler, I., Wolter, F., Junkermann, J.: Sensitive questions in surveys. A comprehensive meta-analysis of experimental survey studies on the performance of the item count technique. Public Opin. Q. **85**(1), 6–27 (2021). https://doi.org/10.1093/poq/nfab002
11. Frye, T., Gehlbach, S., Marquardt, K.L., Reuter, O.J.: Is Putin's popularity real? Post-Soviet Affairs **33**(1), 1–15 (2017). https://doi.org/10.1080/1060586X.2016.1144334
12. Gibson, M.A., Gurmu, E., Cobo, B., Rueda, M.M., Scott, I.M.: Indirect questioning method reveals hidden support for female genital cutting in South Central Ethiopia. PLoS ONE **13**(5), e0193985 (2018). https://doi.org/10.1371/journal.pone.019398
13. Glynn, A.N.: What can we learn with statistical truth serum? Public Opin. Q. **77**(Special Issue), 159–172 (2013). https://doi.org/10.1093/poq/nfs070
14. Groenitz, H.: Improvements and extensions of the item count technique. Electron. J. Stat. **8**, 2321–2351 (2014). https://doi.org/10.1214/14-EJS951

15. Ibrahim, F.: An alternative modified item count technique in sampling survey. Int. J. Stat. Appl. **6**(3), 177–187 (2016). https://doi.org/10.5923/j.statistics.20160603.11
16. Kiewiet de Jonge, C.P., Nickerson, D.W.: Assessing the item count technique in comparative surveys. Polit. Behav. **36**(3), 659–682 (2014). https://doi.org/10.1007/s11109-013-9249-x
17. Malesky, E.J., Gueorguiev, D.D., Jensen, N.M.: Monopoly money: foreign investment and bribery in Vietnam, a Survey Experiment. Am. J. Polit. Sci. **59**(2), 419–439 (2015). https://doi.org/10.1111/ajps.12126
18. Manoli, E.: Advancements in indirect questioning designs. University of Cyprus. Unpublished dissertation (2022)
19. Perri, P.F., Rueda Garcia, M.D.M., Cobo Rodriguez, B.: Multiple sensitive estimation and optimal sample size allocation in the item sum technique. Biomet. J. **60**, 155–173 (2018). https://doi.org/10.1002/bimj.201700021
20. Petroczi, A., Nepusz, T., Cross, P., Taft, H., Shah, S., Deshmukh, N., Schaffer, J., Shane, M., Adesanwo, C., Barker, J., Naughton, D.P.: New non-randomized model to access the prevalence of discriminating behavior: a pilot study on mephedrone. Subst. Abuse Treat. Prev. Policy **6**(20), 1–18 (2011)
21. Rinken, S., Pasadas-del-Amo, S., Rueda, M., Cobo, B.: No magic bullet: estimating anti-immigrant sentiment and social desirability bias with the item-count technique. Qual. Quant. **55**, 2139–2159 (2021). https://doi.org/10.1007/s11135-021-01098-7
22. Särndal, C.-E., Lundström, S.: Estimation in Surveys with Nonresponse. Wiley, Chichester (2005)
23. Särndal, C.-E., Swensson, B., Wretman, J.: Model Assisted Survey Sampling. Springer, New York (1992)
24. Shaw, P.: Estimating a finite population proportion bearing a sensitive attribute from a single probability sample by item count technique. In: Chaudhuri, A., Christofides, T.C., Rao, C.R. (eds.) Handbook of Statistics, Vol. 34 Data Gathering, Analysis and Protection of Privacy Through Randomized Response Techniques: Qualitative and Quantitative Human Traits. Elsevier, Amsterdam (2016)
25. Tian, G.-L., Tang, M.-L., Wu, Q., Liu, Y.: Poisson and negative binomial item count techniques for surveys with sensitive question. Stat. Methods Med. Res. **26**(2), 931–947 (2017). https://doi.org/10.1177/0962280214563345
26. Tourangeau, R., Yan, T.: Sensitive questions in surveys. Psychol. Bull. **133**(5), 859–883 (2007). https://doi.org/10.1037/0033-2909.133.5.859
27. Tsuchiya, T., Hirai, Y., Ono, S.: A study of the properties of the item count technique. Public Opin. Q. **71**(2), 253–272 (2007). https://doi.org/10.1093/poq/nfm012
28. Wolter, F., Laier, B.: The effectiveness of the item count technique in eliciting answers to sensitive questions. An evaluation in the context of self-reported delinquency. Surv. Res. Methods **8**(3), 153–168 (2014)
29. Wolter, F., Mayerl, J., Andersen, H.K., Wieland, T., Junkermann, J.: Overestimation of COVID-19 vaccination coverage in population surveys due to social desirability bias: results of an experimental methods study in Germany. Socius Sociol. Res. Dyn. World **8**, 1–8 (2022). https://doi.org/10.1177/23780231221094749

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.