Electronic Journal of Statistics
Vol. 8 (2014) 2321–2351
ISSN: 1935-7524
DOI: 10.1214/14-EJS951

Improvements and extensions of the item count technique

Heiko Groenitz

Philipps-University Marburg, Department for Statistics (Faculty 02)
Universitätsstraße 25, 35032 Marburg, Germany
e-mail: groenitz@staff.uni-marburg.de

Abstract: The item count technique (ICT) is a helpful tool to conduct surveys on sensitive characteristics such as tax evasion, corruption, insurance fraud, social fraud or drug consumption. The ICT yields cooperation of the respondents by protecting their privacy. There have been several interesting developments on the ICT in recent years. However, some approaches are incomplete while some research questions can not be tackled by the ICT so far. For these reasons, we broaden the existing literature in two main directions. First, we generalize the single sample count (SSC) technique, which is a simplified version of the original ICT, and derive an admissible estimate for the proportion of persons bearing a stigmatizing attribute, bootstrap variance estimates and bootstrap confidence intervals. Moreover, we present both a Bayesian and a covariate extension of the generalized SSC technique. The Bayesian set up allows the incorporation of prior information (e.g., available from a previous study) into the estimation and thus can lead to more efficient estimates. Our covariate extension is useful to conduct regression analysis, i.e., to estimate the effects of explanatory variables on the sensitive characteristic. Second, we establish a new ICT that is applicable to multicategorical sensitive variables such as the number of times a respondent has evaded taxes or the amount of money earned by undeclared work (recorded in classes). The estimation of the distribution of such attributes was not at all treated in the literature on the ICT so far. Therefore, we derive estimates for the marginal distribution of the sensitive characteristic, Bayesian estimates and regression estimates corresponding to our multicategorical ICT.

Keywords and phrases: Sensitive question, socially desired answer, randomized response, missing data problem, Bayesian inference, logistic regression.

Received June 2014.

1. Introduction

The item count technique (ICT) is a method to elicit truthful answers from respondents in surveys on sensitive topics. The basic idea of the ICT, which was originally proposed in Miller (1984, [14]) is as follows. The interviewees are not requested to answer a sensitive question such as “Have you ever evaded taxes?” directly. Instead they receive a list consisting of the sensitive question and some inquiries on nonsensitive items, e.g., “Is your birthday in the first half of the year?”, “Do you have two or more siblings?” or “Is your telephone number odd?", and are instructed to report only the total number of "yes" answers. Replies to individual questions are not revealed. This scheme protects the interviewees' privacy and yields increased cooperation compared with direct questioning. In particular, answer refusal and untruthful socially desired responses are reduced.

The ICT approach has already been applied in several fields. For example, studies on drug use, theft by employees, shoplifting, buying stolen goods, attitudes towards immigrants, racism, undeclared work, voter turnout, and eating disorder are available in the literature. For a detailed list of articles containing concrete studies conducted with the ICT, we refer, for instance, to Tian and Tang (2014, p. 12, [19]) and Blair and Imai (2012, Section 1, [2]).

To estimate the proportion of persons in the population having the sensitive attribute (e.g., having evaded taxes) from ICT data, the so-called difference-in-means estimator is applied in many articles. This estimator possesses a simple representation, however, it may fall out the interval $[0, 1]$. Tsuchiya (2005, [21]) considers a discrete onedimensional covariate and derives estimators for the proportion of persons having a sensitive outcome among the persons possessing a certain value of the covariate. Imai (2011, [12]) describes regression analysis for the ICT. In particular, the author allows arbitrary covariates and derives a nonlinear least square estimate and a maximum likelihood (ML) estimate (MLE). As a specific feature, the estimations in Imai (2011, [12]) involve a certain model for the number of affirmative answers to the nonsensitive questions. Blair and Imai (2012, [2]) build upon Imai (2011, [12]) and develop methods to estimate the social desirability bias as function of the covariates, to tackle multiple sensitive questions, to improve the efficiency, and to detect and correct failures of the ICT. The work of Imai (2011, [12]) is also the fundament for Kuha and Jackson (2014, [13]), who propose a faster algorithm for the ML estimation that additionally delivers an asymptotic variance estimation automatically. Moreover, they suggest further possible specifications for a model regarding the nonsensitive questions. A version of the ICT that is suitable to estimate the mean of a quantitative sensitive characteristic can be found in Chaudhuri and Christofides (2013, [5]) and Trappmann et al. (2014, [20]), where the latter work includes a real-data study whose interviews were conducted already in 2010. The estimation methods in the articles mentioned above demand to divide the respondents in two groups, a control and a treatment group. Here, the respondents in the treatment group contribute information on the sensitive characteristic whereas persons in the control group provide only information on the nonsensitive items. Regarding this, Petroczi et al. (2011, [15]) describe a version of the ICT (the so-called single sample count (SSC) technique) that gets along without control group and can be applied when the distributions of the nonsensitive items are known.

Despite the interesting developments in recent years, the methodological instruments for the ICT still need important extensions and improvements. For instance, the SSC approach by Petroczi et al. (2011, [15]) focuses on the case of four nonsensitive questions where each nonsensitive characteristic possesses a $Bernoulli(1/2)$ distribution. Moreover, they derive an estimator for the proportion of persons bearing the stigmatizing attribute that can attain inadmissible values outside $[0, 1]$. These practical problems motivate us to enhance the work of Petroczi et al. (2011, [15]) by dealing with an arbitrary number of innocuous items whose distributions are not restricted to the $Bernoulli(1/2)$ case, and to develop a feasible estimator in $[0, 1]$. Here, we show that the occurring data situation corresponds to a special missing data pattern and apply the expectation maximization (EM) algorithm to obtain the valid estimator. We establish bootstrap variance estimates for our estimator as well as bootstrap confidence intervals. The bootstrap concept is appropriate not only for large samples, but also for smaller samples. After illustrating our estimation procedure in a real-data example, we demonstrate the efficiency gains that can be realized by the ICT without control group. Furthermore, we derive both a Bayesian and a covariate extension for the generalized SSC procedure. The Bayesian extension is motivated by the fact that sometimes prior information (e.g., from a previous study) is available and should be incorporated into the estimation. The covariate set up enables the researcher to study the dependence of the sensitive variable on nonsensitive exogenous quantities. Our covariate extension is beneficial, for instance, to analyze the effects of gender, nationality, regular occupation, and expected sanctions on conducting undeclared work. Such an analysis on undeclared work delivers an impression of the economic loss caused by moonlighting and may be relevant for economic policy decisions. To illustrate our method for the estimation of the influence of explanatory variables on the sensitive attribute, a numerical example is given in the paper.

Another problem that has not been addressed in the literature on the ICT so far is the estimation of the distribution of multichotomous sensitive characteristics. An example of such variables is income (divided in classes). Income is often relevant in social surveys such as the German General Social Survey (ALLBUS) or the Socio-Economic Panel (SOEP). Nevertheless, when asking directly for income, a large amount of missing values or untruthful answers typically occurs, because persons with higher income are often afraid of envy while persons with lower income are often ashamed. Further examples for multicategorical sensitive characteristics are the number how often one has conducted insurance fraud, the number of hours per week somebody conducts moonlighting, and the monthly amount of income earned by undeclared work (e.g., with categories 0 Euro, 1–100 Euro, 101–1000 Euro, more than 1000 Euro). The consideration of the latter variable is more interesting than considering only the variable describing whether a person conducts undeclared work or not, because a researcher can obtain a more precise impression of the loss through moonlighting.

To fill this gap in the literature, we propose an extension of the ICT to polychotomous sensitive attributes with an arbitrary number of categories in the second part of this paper. In this context, we derive estimates for the unconditional distribution of the sensitive variable, different Bayesian estimates that enable the exploitation of prior knowledge, and regression estimates that are useful for the investigation of the influence of nonsensitive explanatory variables on the polychotomous sensitive quantity.

The paper continues with a review of the ICT by Miller (1984, [14]) in Section 2. In Section 3, we present extensions and improvements of the ICT according to Petroczi et al. (2011, [15]). In Section 4, we establish an ICT for polychotomous sensitive variables. Finally, concluding remarks are available in Section 5.

2. Miller's item count technique

The item count technique according to Miller (1984, [14]) is suitable to gather data on a binary sensitive characteristic. Here, the respondents are randomly divided into a control group and a treatment group. The respondents in the control group receive a list with J nonsensitive questions and have to reveal the number how often they would have to give a "yes" response, i.e., they reply a number between 0 and J. In the treatment group, a list consisting of the same nonsensitive questions and a sensitive question is presented to the interviewees, who have to provide the total number of affirmative answers to these $J+1$ questions, i.e., a number between 0 and $J+1$ must be written in the questionnaire or told to the interviewer.

Formally, let $U_j \in \{0,1\}$ $(j = 1,\dots,J)$ and $Y \in \{0,1\}$ be a nonsensitive and sensitive attribute, respectively. E.g., U_1 and U_2 may indicate whether a person went to a sporting event in the last year and has an even telephone number, respectively. Regarding Y, the value 1 typically represents a stigmatizing attribute (e.g., person has evaded taxes) whereas the value 0 stands for the corresponding nonstigmatizing inverse (person has never evaded taxes). Define $T = 0$ if a person is assigned to the control group and $T = 1$ if a person belongs to the treatment group. Moreover, set $Z = U_1 + \dots + U_J$. Then, the required answer S of a person in the control group is Z while interviewees in the treatment group are instructed to give an answer $Z + Y$. In the control group, nobody is confronted with any sensitive item so that truthful answers can be supposed. In most cases, the privacy of the persons in the treatment group is protected and truthful answers can be expected, because only a total and not the value of Y is reported. Notice, however, that the protection of the privacy can fail when all nonsensitive items apply (i.e., $U_1 = \dots = U_J = 1$). In this case, an answer $J + 1$ implies $Y = 1$. To minimize this "ceiling effect", one should select nonsensitive questions for which only few persons would give throughout "yes" answers. Furthermore, $Y = 0$ follows from an answer 0. However, if $Y = 0$ represents a nonstigmatizing outcome (e.g., no tax evasion), this "floor effect" is less problematic than the ceiling effect.

Let us assume that a simple random sample of n persons has been drawn and denote the ith sample unit's outcome corresponding to U_j, Y, T, Z, S by U_{ij}, Y_i, T_i, Z_i, S_i, respectively. Further, denote the proportion of persons in the universe having $Y = 1$ by π_1, set $\pi_0 = 1 - \pi_1$, and define $\pi = (\pi_0, \pi_1)^\top$. To estimate π_1, the difference-in-means estimator

$$\hat{\pi}_1 = n_T^{-1} \sum_{i=1}^{n} T_i S_i \;-\; n_C^{-1} \sum_{i=1}^{n} (1 - T_i) S_i \tag{1}$$

with $n_T = \sum_{i=1}^{n} T_i$ and $n_C = n - n_T$ is used in many articles. Unfortunately, this estimator can attain negative values and values greater than one. We remark that an ICT with a slightly different procedure and a corresponding estimator are proposed by Chaudhuri and Christofides (2007, [4]). This modified version of the ICT avoids a ceiling effect, but also inheres floor effects. Moreover, the estimator in Chaudhuri and Christofides (2007, [4]) can attain values outside $[0, 1]$, too.

3. Improvements and Extensions of the Single Sample Count Technique

In an article by Petroczi et al. (2011, [15]) on a study on Mephedrone use, a variant of the ICT without control group is mentioned. Petroczi et al. (2011, [15]) call their version of the ICT the single sample count technique. However, these authors focus on the case of $J = 4$ nonsensitive items, assume that the distribution of U_j $(j = 1, \ldots, 4)$ is known and equal to a Bernoulli distribution with probability of success 1/2, and suggest a moment-based estimator for π_1 that can fall out the interval $[0, 1]$. These practical limitations motivate us to extend the approach by Petroczi et al. (2011, [15]) and develop admissible estimates between 0 and 1 for the proportion π_1. Moreover, we develop Bayesian estimates and present a method that enables regression analysis, i.e., the investigation of the influence of covariates on the sensitive item.

3.1. General procedure and ML estimation

Let us consider the following general procedure for an ICT without control group. Each interviewee in the sample is supplied with a list of J ($J \in \mathbb{N}$ arbitrary) nonsensitive questions supplemented by a question on a sensitive topic and is instructed to reveal only the total number of affirmative answers. Continuing the notation from Section 2, each respondent gives the answer $S = Z + Y \in \{0, \ldots, J+1\}$. Compared with Section 2, we now only have a treatment group and every respondent contributes information on the distribution of Y. Regarding the protection of the privacy, we have analog statements as in Section 2. In particular, the answers $1, \ldots, J$ do not expose the respondent's Y-value. However, answer 0 implies $Y = 0$. Even more problematic is that $Y = 1$ can be concluded from response $J+1$. The risk that a truthfully answering interviewee must admit that he or she possesses $Y = 1$ (i.e., must give answer $J+1$) can be reduced by selecting control items such that the proportion of individuals in the population having $Z = J$ is small. An alternative approach that improves the protection of the privacy is sketched in Section 5. We proceed with two assumptions:

$$\text{the distribution of } Z \text{ is known and} \tag{2}$$

$$Z \text{ and } Y \text{ are independent.} \tag{3}$$

We discuss these assumptions in Subsection 3.4. To estimate the marginal distribution of Y, we propose maximum likelihood estimation rather than moment estimation, because the MLE for π_1 is always admissible, i.e., in $[0,1]$. To compute the MLE, the EM algorithm due to Dempster, Laird, and Rubin (1977, [7]) is beneficial. Hereto, note that $\mathbf{S} = (S_1, \dots, S_n)$ describes our observed data while $\mathbf{Y} = (Y_1, \dots, Y_n)$ and $\mathbf{Z} = (Z_1, \dots, Z_n)$ are missing values. We denote the proportion of individuals in the population having $Z = i$ with ϕ_i $(i = 0, \dots, J)$ and set $\phi = (\phi_0, \dots, \phi_J)^\top$, that is, ϕ is known due to (2). Further, set $\lambda = (\lambda_0, \dots, \lambda_{J+1})^\top$ where λ_i is the proportion of units in the population possessing $S = i$ and assume that the n sample units were drawn by simple random sampling with replacement (SRSWR). Let $\mathbf{s} = (s_1, \dots, s_n)$, $\mathbf{y} = (y_1, \dots, y_n)$, and $\mathbf{z} = (z_1, \dots, z_n)$ be the realizations of $\mathbf{S}$, $\mathbf{Y}$, and $\mathbf{Z}$, respectively. The observed data log-likelihood is given by

$$l_{obs}(\pi; \mathbf{s}) = \sum_{i=1}^{n} \log \mathbb{P}(S_i = s_i) = \sum_{i=1}^{n} \log \left[\phi_{s_i} \cdot \pi_0 + \phi_{s_i - 1} \cdot \pi_1\right]$$

with the convention $\phi_x = 0$ if $x \notin \{0, \dots, J\}$. Similar conventions are used permanently in the paper, either explicitly or implicitly. For the complete data log-likelihood,

$$l_{com}(\pi) = l_{com}(\pi; \mathbf{y}, \mathbf{z}, \mathbf{s}) = \sum_{i=1}^{n} \log \mathbb{P}(Y_i = y_i, Z_i = z_i, S_i = s_i) =$$

$$\sum_{i=1}^{n} \log \mathbb{P}(Y_i = y_i) + const. = \log \pi_0 \cdot \sum_{i=1}^{n} 1_{\{0\}}(y_i) + \log \pi_1 \cdot \sum_{i=1}^{n} 1_{\{1\}}(y_i) + const.$$

holds. Applying the EM algorithm to maximize l_{obs}, each iteration consists of an E step and a M step. When $\pi^{(t)} = (\pi_0^{(t)}, \pi_1^{(t)})^\top$ is available from the preceding iteration t, we calculate an estimated complete data log-likelihood in the E step of iteration $t+1$ by

$$\begin{aligned}
\widehat{l_{com}}(\pi) \quad &= \quad \mathbb{E}_t(l_{com}(\pi; \mathbf{Y}, \mathbf{Z}, \mathbf{S}) \,|\, \mathbf{S} = \mathbf{s}) = \log \pi_0 \cdot \sum_{i=1}^{n} \mathbb{E}_t(1_{\{0\}}(Y_i) | \mathbf{S} = \mathbf{s}) \\
&+ \quad \log \pi_1 \cdot \sum_{i=1}^{n} \mathbb{E}_t(1_{\{1\}}(Y_i) | \mathbf{S} = \mathbf{s}) + const. \\
&=: \quad \log \pi_0 \cdot v_0^{(t)} + \log \pi_1 \cdot v_1^{(t)} + const.
\end{aligned} \tag{4}$$

where $\mathbb{E}_t$ and $\mathbb{P}_t$ (see below) mean the calculation of expectation and probability assuming $\pi^{(t)}$ is the true parameter. We can further compute the expectations

$$\mathbb{E}_t(1_{\{j\}}(Y_i) | \mathbf{S} = \mathbf{s}) = \mathbb{P}_t(Y_i = j | S_i = s_i) = \frac{\phi_{s_i - j} \cdot \pi_j^{(t)}}{\phi_{s_i} \cdot \pi_0^{(t)} + \phi_{s_i - 1} \cdot \pi_1^{(t)}} \quad (j = 0, 1).$$

Notice, we have the compact representation

$$\begin{pmatrix} v_0^{(t)} \\ v_1^{(t)} \end{pmatrix} = \left(\boldsymbol{\phi} \,.^* \left[\begin{pmatrix} 1/\lambda_0^{(t)} \\ \vdots \\ 1/\lambda_{J+1}^{(t)} \end{pmatrix} \cdot (\pi_0^{(t)}, \pi_1^{(t)}) \right] \right)^\top \cdot n_1^\top =: P^{(t)} \cdot n_1^\top. \tag{5}$$

Here, the entry (i, j) of the $2 \times (J+2)$ matrix $P^{(t)}$ is equal to $\mathbb{P}_t(Y = i|S = j)$ $(i = 0, 1;\ j = 0, \ldots, J+1)$, the entry (i, j) of the $(J+2) \times 2$ matrix $\boldsymbol{\phi}$ equals ϕ_{i-j} $(i = 0, \ldots, J+1;\ j = 0, 1)$, $\lambda^{(t)} = (\lambda_0^{(t)}, \ldots, \lambda_{J+1}^{(t)})^\top = \boldsymbol{\phi} \cdot \pi^{(t)}$, $n_1 = (n_{10}, \ldots, n_{1,J+1})$ where n_{1i} equals the number how often answer i occurred in the sample, and $.^*$ denotes componentwise multiplication. The maximum of the function $\widehat{l_{com}}$, which is calculated in the M step of iteration $t+1$, is given by

$$\pi_0^{(t+1)} = \frac{v_0^{(t)}}{v_0^{(t)} + v_1^{(t)}}, \quad \pi_1^{(t+1)} = \frac{v_1^{(t)}}{v_0^{(t)} + v_1^{(t)}}.$$

After choosing a starting value, e.g., $\pi^{(0)} = (0.5,\, 0.5)^\top$, we obtain step-by-step a sequence $(\pi^{(t)})_{t \in \mathbb{N}_0}$, for which the corresponding values of the observed data log-likelihood are nondecreasing. When the variation from $\pi^{(t)}$ to $\pi^{(t+1)}$ is small enough, we have found an estimate $\hat{\pi}$.

Since we have no handy analytic representation of $\hat{\pi}$, the bootstrap (BS) approach is attractive for the computation of standard errors and confidence intervals (CIs). Here, we calculate B bootstrap replications of $\hat{\pi}$, denoted by $\hat{\pi}^{(b)}$ for $b = 1, \ldots, B$. The empirical variance of these replications is the BS estimate $\widehat{Var}_{BS}(\hat{\pi})$ for the variance of $\hat{\pi}$. The square roots of the diagonal elements of $\widehat{Var}_{BS}(\hat{\pi})$ represent the BS standard errors of the components of $\hat{\pi}$. The empirical $\alpha/2$ quantile of the replications of the ith component of $\hat{\pi}$ provides a lower bound of a $1 - \alpha$ CI for π_i while an upper bound is given by the $1 - \alpha/2$ quantile. To obtain one replication $\hat{\pi}^{(b)}$, we treat $\hat{\pi} = (\hat{\pi}_0, \hat{\pi}_1)^\top$ as true parameter and simulate new frequencies of the answers $0, \ldots, J+1$ by

$$n_1^{(b)} = (n_{10}^{(b)}, \ldots, n_{1,J+1}^{(b)}) \sim Multinomial(n, (\hat{\lambda}_0, \ldots, \hat{\lambda}_{J+1})) \tag{6}$$

where $\hat{\lambda}_i = \phi_i \hat{\pi}_0 + \phi_{i-1} \hat{\pi}_1$. The quantity $\hat{\pi}^{(b)}$ is then obtained by applying the EM algorithm to the new frequencies $n_1^{(b)}$.

3.2. Real-data example

One may think that the EM algorithm is not necessary, because $\hat{\pi}_1 = 0$ could be set if the moment estimate according to Petroczi et al. (2011, [15]) is negative. Then, however, $\hat{\pi}_1 = 0$ is in general not the MLE. To show this, we reanalyze data from Petroczi et al. (2011, [15]) where Y equals 1 if a person has taken the drug Mephedrone at least once in the previous three months, and Y equals 0 else. Moreover, Z can attain the values 0, 1, 2, 3, and 4 and follows a

$Binomial(4, 0.5)$ distribution. Consequently, the possible answers S are 0, 1, 2, 3, 4, and 5. According to Petroczi et al. (2011, Table 3, [15]), the answer 0, 1, 2, 3, 4, and 5 was observed 15, 64, 89, 51, 16, and 2 times, respectively. Petroczi et al. (2011, [15]) calculate the negative moment estimate -0.0211 for π_1 and an asymptotic 95% confidence interval for π_1 equal to [0, 0.0996].

Since this moment estimate is inappropriate, we calculate the MLE for π_1 via EM algorithm. With starting value $\pi_1^{(0)} = 0.5$, we obtain the MLE $\hat{\pi}_1 = 0.0632$ within 58 iterations. This example underlines that the EM algorithm is beneficial to compute the desired valid estimate. Additionally, we obtain 0.0496 as the bootstrap standard error of $\hat{\pi}_1$ and [0, 0.1671] as a 95% bootstrap CI for π_1.

3.3. Increase of accuracy

We now demonstrate the efficiency gains that can be achieved by using control items with known distribution and dispensing with the control group. For this purpose, we compare two procedures. Procedure one is the ICT without control group from Subsection 3.1, for which every respondent contributes information on Y and π can be estimated via EM algorithm as shown. The second procedure is the ICT with control group according to Section 2 where we assume that every sample unit is assigned to the control group with probability 50%. For the second procedure, we can apply an EM algorithm for the ML estimation, too. This is implicitly contained in Imai (2011, [12]). Moreover, this estimation is a special case of the estimation in Subsection 4.1 (set $k = 2$, $k_1 = \cdots = k_J = 1$ in Subsection 4.1). Our comparison is conducted by some simulations, in which we consider $\pi = (0.8, 0.2)^\top$ and three specifications of ϕ resulting in the cases I–III given in Table 1. In procedure 1, ϕ is known and in procedure 2, ϕ is unknown.

For each case and each procedure (ICT without or with control group), we simulate 10000 samples with sample size 250. For each sample, we calculate the corresponding estimate for π. The 10000 generated realizations of an estimator are used to compute the simulated expectations and MSEs of the estimator's components. The results are given in Table 2. We recognize that the simulated bias of each estimator is close to zero. Moreover, the application of control items whose sum has a known distribution leads to manifestly more efficient estimates. This is not surprising, because we can expect that the incorporation of more information (here, the additional information is the known distribution of Z) results in better estimates.

TABLE 1

The specifications of ϕ, which represents the distribution of Z. Each ϕ is obtained by determining the marginal distributions of U_i and assuming independence of the U_i. $Ber(p)$ means a Bernoulli distribution with parameter p. The distribution of the sensitive Y is always $\pi = (0.8, 0.2)^\top$

case	U_1	U_2	U_3	U_4	$\phi^\top$				
I	$Ber(0.5)$	$Ber(0.5)$	$Ber(0.5)$	$Ber(0.5)$	0.0625	0.2500	0.3750	0.2500	0.0625
II	$Ber(0.2)$	$Ber(0.2)$	$Ber(0.5)$	$Ber(0.5)$	0.1600	0.4000	0.3300	0.1000	0.0100
III	$Ber(0.2)$	$Ber(0.2)$	$Ber(0.4)$	$Ber(0.5)$	0.1920	0.4160	0.3000	0.0840	0.0080

TABLE 2
Simulation results for the comparison between the ICT from Subsection 3.1 and an ICT with control group

	ICT without control group				ICT with control group			
case	$\hat{\mathbb{E}}(\hat{\pi}_0)$	$\hat{\mathbb{E}}(\hat{\pi}_1)$	$M\hat{S}E(\hat{\pi}_0)$	$M\hat{S}E(\hat{\pi}_1)$	$\hat{\mathbb{E}}(\hat{\pi}_0)$	$\hat{\mathbb{E}}(\hat{\pi}_1)$	$M\hat{S}E(\hat{\pi}_0)$	$M\hat{S}E(\hat{\pi}_1)$
I	0.8008	0.1992	0.0040	0.0040	0.7919	0.2081	0.0181	0.0181
II	0.8009	0.1991	0.0037	0.0037	0.7966	0.2034	0.0140	0.0140
III	0.7999	0.2001	0.0038	0.0038	0.7969	0.2031	0.0143	0.0143

3.4. Discussion of assumptions

Potential control items can be constructed in various ways, some examples are: “Is your birthday in the first quarter of the year?”, “Is the last digit of your best friend’s telephone number equal to 7, 8, or 9?”, “Are you born during the dates 1–10?”, “Is your mother’s birthday this year on Monday or Tuesday?”, “Do you have two or more siblings?”, “Are you the owner of a house?”, “Do you have a university degree?”, “Is your house number an even number?”, “Are you evangelic?”.

For some of these control items, there is an intuition about the marginal distribution of the characteristic. E.g., for the first control item, the probability of success can be assumed to be 1/4 while for the second control item, the probability of success should be approximately 3/10. If we have no such intuition or if we would like to use more precise quantities, we can work with census data or other databases (e.g., Petroczi et al. (2011, p. 9–10 and p. 15–16, [15]) mention some data sources for house and phone numbers as well as birthdays). For above control items independence between the variables is often a reasonable assumption. When we have independence and know the marginal distributions, the distribution of Z can be calculated. Notice, if we do not want to assume independence between some control items, we sometimes are able to find their precise joint distribution e.g. in census data. All in all, an investigator will usually find control items for which it is justifiable to assume (2) and (3). Regarding the number of control items, we first note that the application of a very small number of control items (say, 1–2) is not recommendable from the viewpoint of privacy protection. For too many control items (e.g., 10 or more), the efficiency of the corresponding estimate is doubtful. Maybe 4–5 items seem to be an appropriate choice.

We now study the case where (2) is not fulfilled. In this situation, we have to estimate ϕ, too. The ML estimation for $\theta = (\pi^\top, \phi^\top)^\top$ can be conducted via EM algorithm. Because the concrete algorithm needed for this problem is a special case of the EM algorithm for the maximization of (11) in Section 4 (set $k = 2$, $k_1 = \cdots = k_J = 1$, and $t_1 = \cdots = t_n = 1$ in Subsection 4.1), we exclude further details on the iterations here.

However, when we conduct a ML estimation for $\theta = (\pi^\top, \phi^\top)^\top$ for an ICT procedure without control group, identification problems become manifest, i.e., we have nonunique MLEs. Let us consider the cases A-C from Table 3 to construct an illustrative example. In each case, the distributions of the U_i and Y

TABLE 3
Three specifications for U_1, U_2, Y that all result in the same probabilities of the answers 0, 1, 2, 3 (independence of U_1, U_2, Y is assumed). Ber(p) represents a Bernoulli distribution with probability of success equal to p

case	U_1	U_2	Y	$\phi^\top$			$\lambda^\top = (\lambda_0, \ldots, \lambda_3)$			
A	$Ber(0.5)$	$Ber(0.8)$	$Ber(0.2)$	0.10	0.50	0.40	0.08	0.42	0.42	0.08
B	$Ber(0.2)$	$Ber(0.8)$	$Ber(0.5)$	0.16	0.68	0.16	0.08	0.42	0.42	0.08
C	$Ber(0.5)$	$Ber(0.2)$	$Ber(0.8)$	0.40	0.50	0.10	0.08	0.42	0.42	0.08

are specified. Although these specifications are different, they all lead to the same distribution of the answers. Now assume that the absolute frequency of the answer 0, 1, 2, 3 in the sample equals 8, 42, 42, 8, respectively. Then, it is not surprising that each of the vectors

$$(0.8, 0.2, 0.1, 0.5, 0.4)^\top, (0.5, 0.5, 0.16, 0.68, 0.16)^\top, (0.2, 0.8, 0.4, 0.5, 0.1)^\top$$

is a MLE for θ. Owing to this identification problem, it is not recommendable to apply the ICT without control group when ϕ must be estimated from our survey data.

3.5. Bayesian estimation

Sometimes prior information on the distribution of Y is available, e.g., from an earlier study. By incorporating prior information into the estimation, we can expect to obtain better, i.e., more accurate, estimates. Such estimates can be calculated by application of Bayesian methods. In a Bayesian context, the parameter (π_0, π_1) is considered to be a realization of a random vector (Π_0, Π_1). The investigator has to define a distribution of (Π_0, Π_1) (the so called prior distribution) which contains the prior information. For the conditional density of the complete data $(\mathbf{Y}, \mathbf{S})$ given a value of Π_0, we set for $y_i \in \{0, 1\}$ and $s_i \in \{0, \ldots, J+1\}$

$$f_{\mathbf{Y},\mathbf{S}\,|\,\Pi_0}(\mathbf{y}, \mathbf{s}\,|\,\pi_0) = \prod_{i=1}^{n} \boldsymbol{\phi}(s_i, y_i) \cdot \pi_{y_i}, \tag{7}$$

where $\boldsymbol{\phi}(i, j)$ is entry (i, j) of matrix $\boldsymbol{\phi}$ from (5) and $\pi_1 = 1 - \pi_0$. Consequently, (7) and the prior of Π_0 completely determine the distribution of $(\mathbf{Y}, \mathbf{S}, \Pi_0)$. Let us assume an outcome $\mathbf{s}$ of $\mathbf{S}$ has been recorded in our survey with the item count technique without control group. Then, the principle of Bayes inference is to evaluate the posterior distribution of $(\Pi_0, \mathbf{Y})$ given the observed $\mathbf{s}$. This results in estimates that base on both the prior information and the information in $\mathbf{s}$ from the current survey. In the sequel, we give more details on the Bayes estimation for the ICT without control group.

Regarding the prior distribution, we apply the $Beta(\delta_0, \delta_1)$ distribution for Π_0, that is, we assume Π_0 to have density

$$f_{\Pi_0}(\pi_0) = K \cdot \pi_0^{\delta_0 - 1} \cdot (1 - \pi_0)^{d_1 - 1} \cdot 1_{[0,1]}(\pi_0),$$

where $\delta_0, \delta_1 > 0$ are parameters and the constant K depends on the δ_i. Clearly, we have the uniform distribution on $[0, 1]$ for $\delta_0 = \delta_1 = 1$. We consider the Beta distribution, because of the following properties. First, the Beta prior is interpretable well. In particular, the $Beta(\delta_0, \delta_1)$ distribution contains the same information as $(\delta_0 - 1) + (\delta_1 - 1)$ additional observations among which the outcomes $Y = 0$ and $Y = 1$ occur $\delta_0 - 1$ times and $\delta_1 - 1$ times, respectively. Second, an investigator's guess $\hat{\pi}_0^{(p)}$ for π_0, which may be based on a previous study, can be transformed into a concrete Beta prior so that the certainty about the guess is reflected. For this purpose, let us fix a proportionality constant d and set $\delta_0 = \hat{\pi}_0^{(p)} \cdot d$ as well as $\delta_1 = (1 - \hat{\pi}_0^{(p)}) \cdot d$. Then, the Beta prior with these parameters δ_i comprises the same information as $d - 2$ new observations. Thus, a large d corresponds to a large certainty of the investigator about the guess $\hat{\pi}_0^{(p)}$. Third, the Beta prior allows comparatively comfortable calculations for the EM and data augmentation algorithm (see below).

We next derive several possibilities to study the posterior distribution of $(\Pi_0, \mathbf{Y})$ given $\mathbf{s}$. We start with the calculation of the mode of the density $f_{\Pi_0|\mathbf{S}}(\cdot \,|\, \mathbf{s})$. We remark that in the case of a uniform prior, this posterior mode equals the MLE. Dempster, Laird, and Rubin (1977, [7]) show for general missing data constellations that a version of the EM algorithm can be used to detect the posterior mode. In our situation of the ICT without control group, the posterior mode calculation adds up to modify the EM algorithm for the MLE from above. In the E step of iteration $t + 1$, we now compute the function

$$\pi_0 \mapsto \log \pi_0 \cdot v_0^{(t)} + \log(1 - \pi_0) \cdot v_1^{(t)} + \log f_{\Pi_0}(\pi_0) \tag{8}$$

with $v_i^{(t)}$ as in (5). Compared with (4), the term $\log f_{\Pi_0}(\pi_0)$ corresponding to the prior distribution now appears. The maximum of (8) is searched in the M step. It is equal to

$$\pi_0^{(t+1)} = \frac{v_0^{(t)} + \delta_0 - 1}{n + \delta_0 + \delta_1 - 2}.$$

Beginning with a starting value, this EM procedure produces step-by-step a sequence $\pi_0^{(0)}, \pi_0^{(1)}, \pi_0^{(2)}, \ldots$ with $f_{\Pi_0|\mathbf{S}}(\pi_0^{(t+1)} \,|\, \mathbf{s}) \geq f_{\Pi_0|\mathbf{S}}(\pi_0^{(t)} \,|\, \mathbf{s})$ (cf. Schafer (2000, p. 46, [17]) for a general missing data problem).

Further possibilities to evaluate $f_{\Pi_0 \,|\, \mathbf{S}}(\cdot \,|\, \mathbf{s})$ are to calculate the expectation, i.e., the posterior mean, as another point estimate for the true π_0 and quantiles as bounds of confidence intervals. Moreover, we can consider the relative frequency of sample units having the outcome $Y = 0$, i.e., we look at $P_0 = \frac{1}{n} \sum_{i=1}^{n} 1_{\{Y_i = 0\}}$ and compute the expectation and quantiles of the distribution of P_0 given $\mathbf{S} = \mathbf{s}$. To detect the mentioned expectations and quantiles, the data augmentation (DA) algorithm (see Tanner and Wong (1987, [18])) is helpful. With this iterative procedure, we obtain a sequence of realizations $(\mathbf{y}^{(t)}, \pi_0^{(t)})_{t \in \mathbb{N}}$ of a Markov chain $(\mathbf{Y}^{(t)}, \Pi_0^{(t)})_{t \in \mathbb{N}}$, which converges in distribution to the distribution given by the conditional density $f_{\mathbf{Y}, \Pi_0 \,|\, \mathbf{S}}(\cdot, \cdot \,|\, \mathbf{s})$. In particular, in the I step of iteration $t + 1$ of the DA scheme, we must draw a vector

$\mathbf{y}^{(t+1)}$ from $f_{\mathbf{Y}\,|\,\mathbf{S},\Pi_0}(\cdot\,|\,\mathbf{s},\pi_0^{(t)})$. Regarding this, (7) implies

$$f_{\mathbf{Y}\,|\,\mathbf{S},\Pi_0}(\mathbf{y}\,|\,\mathbf{s},\pi_0^{(t)}) = \prod_{i=1}^{n} \frac{\phi(s_i, y_i)\cdot \pi_{y_i}^{(t)}}{f_{S_i\,|\,\Pi_0}(s_i\,|\,\pi_0^{(t)})},$$

where $\pi_1^{(t)} = 1-\pi_0^{(t)}$ and $f_{S_i\,|\,\Pi_0}(s_i\,|\,\pi_0^{(t)})$ equals the entry number $s_i \in \{0,\ldots,J+1\}$ of the vector $\boldsymbol{\phi}\cdot(\pi_0^{(t)},\pi_1^{(t)})^\top$. In the subsequent P step, we generate a new parameter $\pi_0^{(t+1)}$ from the density $f_{\Pi_0\,|\,\mathbf{Y},\mathbf{S}}(\cdot\,|\,\mathbf{y}^{(t+1)},\mathbf{s})$. According to (7) and the $Beta(\delta_0,\delta_1)$ prior, this density corresponds to a $Beta(m_0^{(t+1)}+\delta_0, m_1^{(t+1)}+\delta_1)$ distribution. Here, we define $m_0^{(t+1)} = \sum_{i=1}^n 1_{\{0\}}(y_i^{(t+1)})$ and $m_1^{(t+1)} = n - m_0^{(t+1)}$ where $y_i^{(t+1)}$ is the ith entry of $\mathbf{y}^{(t+1)}$. Due to a strong law of large numbers (SLLN) for Markov chains (for instance, Schafer (2000, p. 91, [17])), we obtain for $L\to\infty$ the almost sure convergences

$$\hat{p}_L = \frac{1}{L}\sum_{t=1}^{L}\Pi_0^{(t)} \xrightarrow{a.s.} \mathbb{E}(\Pi_0\,|\,\mathbf{S}=\mathbf{s}) \text{ and } F_L(x) = \frac{1}{L}\sum_{t=1}^{L} 1_{\{\Pi_0^{(t)}\le x\}} \xrightarrow{a.s.} F_{\Pi_0|\mathbf{S}}(x|\mathbf{s}),$$

where $F_{\Pi_0|\mathbf{S}}(\cdot\,|\,\mathbf{s})$ is the distribution function of Π_0 given $\mathbf{s}$. Accordingly, the quantile functions also converge, that is, for $u\in(0,1)$, we have

$$F_L^{-1}(x) \xrightarrow{a.s.} F^{-1}_{\Pi_0\,|\,\mathbf{S}}(x\,|\,\mathbf{s}) \text{ for } L\to\infty.$$

It is appealing to use the a.s. limit of $\hat{p}_L$ as point estimate for the true π_0 while the a.s. limits of $F_L^{-1}(\alpha/2)$ and $F_L^{-1}(1-\alpha/2)$ provide a lower and an upper bound of a $1-\alpha$ confidence interval for the true proportion π_0. These limits can be simulated with the help of the DA algorithm as described above.

Let us now analyze the distribution of P_0 given $\mathbf{S}=\mathbf{s}$. The values $m_0^{(t)}$ $(t\ge 1)$ can be interpreted as multiple imputations (MIs) for $\sum_{i=1}^n 1_{\{Y_i=0\}}$. Set $P_0^{(t)} = M_0^{(t)}/n$ where $M_0^{(t)}$ is the random variable that belongs to $m_0^{(t)}$ and introduce $\hat{p}_L^{MI} = \frac{1}{L}\sum_{t=1}^L P_0^{(t)}$. Then, the Markov chain SLLN guarantees

$$\hat{p}_L^{MI} \xrightarrow{a.s.} \mathbb{E}(P_0\,|\,\mathbf{S}=\mathbf{s}) \text{ and } F_L^{MI}(x) = \frac{1}{L}\sum_{t=1}^{L} 1_{\{P_0^{(t)}\le x\}} \xrightarrow{a.s.} F_{P_0|\mathbf{S}}(x|\mathbf{s}),$$

where $F_{P_0|\mathbf{S}}(\cdot\,|\,\mathbf{s})$ is the distribution function of P_0 given $\mathbf{s}$. When Q_L^{MI} represents the quantile function that belongs to F_L^{MI}, it follows that $Q_L^{MI}(u) \xrightarrow{a.s.} F^{-1}_{P_0|\mathbf{S}}(u|\mathbf{s})$ for any $u\in(0,1)$ where the quantile function $F^{-1}_{P_0|\mathbf{S}}(\cdot,\mathbf{s})$ is continuous. The a.s. limit of $\hat{p}_L^{MI}$ is another point estimate for the true π_0 and the a.s. limits of $Q_L^{MI}(\alpha/2)$ and $Q_L^{MI}(1-\alpha/2)$ deliver CI bounds. These limits can be detected by DA, too.

We close this subsection with the remark that the Bayesian estimation methods established above concretely address the item count technique without control group. However, for various randomized response and nonrandomized response procedures, Bayesian estimates can be derived in a similar way. In this regard, the interested reader is referred to Groenitz (2013, [9]).

3.6. Covariate extension

In this subsection, we present a covariate extension of the ICT without control group, that is, we develop a method that enables the analysis of the influence of a vector of p nonsensitive covariates X on the sensitive characteristic Y. Such a technique is helpful, for instance, for the investigation of the dependence of tax evasion on gender, age, and profession. We again make the assumption that the distribution of Z in the population is known. We start with the case of deterministic exogenous variables. Here, the researcher determines values of the covariates. Subsequently, persons with these covariate values are randomly selected and requested to give an answer according to the ICT with control group, i.e., each person should reply his or her outcome of $S = Z + Y$. Let x_{ij} be the ith interviewee's value of the jth covariate, and set $x_i = (x_{i1}, \dots, x_{ip})$. We further assume:

(D1) $Y_1, \dots, Y_n$ are independent.
(D2) $Z_1, \dots, Z_n$ are independent and identically distributed (iid) with $Z_i \sim Z$.
(D3) The vectors $(Y_1, \dots, Y_n)$ and $(Z_1, \dots, Z_n)$ are independent.
(D4) There is a $\beta \in \mathbb{R}^p$ with $\mathbb{P}(Y_i = 1) = \frac{e^{x_i\beta}}{1+e^{x_i\beta}}$ $(i = 1, \dots, n)$.

(D1)–(D3) are fulfilled if (Y, X) and Z are independent and if for each covariate level fixed by the researcher, the interviewees are selected by SRSWR from the population units possessing this covariate level where the selection for one covariate level is independent of the selection for the other covariate levels. The assumptions (D1)–(D4) mean that a logistic regression model for the dependence of the sensitive item on the exogenous characteristics holds. To estimate β via EM algorithm, we initially notice the observed data log-likelihood

$$\begin{aligned} l_{obs}(\beta) &= \sum_{i=1}^{n} \log \mathbb{P}(S_i = s_i) = \sum_{i=1}^{n} \log \left[\phi_{s_i} \cdot \frac{1}{1+e^{x_i\beta}} + \phi_{s_i - 1} \cdot \frac{e^{x_i\beta}}{1+e^{x_i\beta}} \right] \\ &= \sum_{r=1}^{R} \sum_{j=0}^{J+1} n_1(r,j) \cdot \log \left[\phi_j \cdot \frac{1}{1+e^{x_{i_r}\beta}} + \phi_{j-1} \cdot \frac{e^{x_{i_r}\beta}}{1+e^{x_{i_r}\beta}} \right]. \end{aligned} \tag{9}$$

Regarding this equality, we assume that $R \leq n$ different covariate levels are in place, that sample unit number i_r possesses the rth covariate level, and that $n_1(r, j)$ is the number how often answer j occurred among the interviewees with the rth covariate level. We have introduced the quantity R to hint that the number of computations and hence the elapsed time of the algorithm can be reduced if the number of different covariate levels is clearly smaller than n. The complete data log-likelihood is apart from a constant equal to

$$l_{com}(\beta) = \sum_{i=1}^{n} \left(1_{\{1\}}(y_i) \cdot \log \frac{e^{x_i\beta}}{1+e^{x_i\beta}} + 1_{\{0\}}(y_i) \cdot \log \frac{1}{1+e^{x_i\beta}} \right).$$

In the E step of iteration $t+1$ of the EM algorithm, we obtain an estimated complete data log-likelihood

$$\widehat{l_{com}}(\beta) = \sum_{i=1}^{n}\left(\mathbb{P}_t(Y_i=1|S_i=s_i)\log\frac{e^{x_i\beta}}{1+e^{x_i\beta}} + \mathbb{P}_t(Y_i=0|S_i=s_i)\log\frac{1}{1+e^{x_i\beta}}\right)$$

$$= \sum_{r=1}^{R}\sum_{j=0}^{J+1} n_1(r,j)\left[\mathbb{P}_t(Y_{i_r}=1|S_{i_r}=j)\log\frac{e^{x_{i_r}\beta}}{1+e^{x_{i_r}\beta}} + \mathbb{P}_t(Y_{i_r}=0|S_{i_r}=j)\log\frac{1}{1+e^{x_{i_r}\beta}}\right]$$

where

$$\mathbb{P}_t(Y_{i_r}=1|S_{i_r}=j) = \frac{\phi_{j-1}\cdot e^{x_{i_r}\beta_{(t)}}}{\phi_{j-1}\cdot e^{x_{i_r}\beta_{(t)}} + \phi_j}$$

holds where $\beta_{(t)}$ is the estimate corresponding to the preceding iteration. In the subsequent M step, we compute a new estimate $\beta_{(t+1)}$. This $\beta_{(t+1)}$ equals the MLE corresponding to a logistic regression model with data such that for covariate level r, $Y=1$ occurs $\sum_{j=0}^{J+1} n_1(r,j)\cdot\mathbb{P}_t(Y_{i_r}=1|S_{i_r}=j)$ times and $Y=0$ occurs $\sum_{j=0}^{J+1} n_1(r,j)\cdot\mathbb{P}_t(Y_{i_r}=0|S_{i_r}=j)$ times (noninteger numbers may appear). Such an MLE can be obtained by standard software (e.g., in MATLAB, one may apply the function `mnrfit`). When the difference between $\beta_{(t)}$ and $\beta_{(t+1)}$ is sufficiently small, we stop iterations and use the last $\beta_{(t)}$ as estimate $\hat{\beta}$.

Estimated standard errors for the components of $\hat{\beta}$ can be obtained via the bootstrap approach. Here, replications $\hat{\beta}_{(1)},\ldots,\hat{\beta}_{(B)}$ are computed and the empirical variance of these replications is the BS estimate for the variance of $\hat{\beta}$. Taking the square roots of the diagonal entries of this matrix yields BS standard errors for the components of $\hat{\beta}$. To generate $\hat{\beta}_{(b)}$ $(b=1,\ldots,B)$, we draw a replication for each $n_1(r,j)$ by

$$(n_1^{(b)}(r,0),\ldots,n_1^{(b)}(r,J+1)) \sim \mathit{Multinomial}(n_r,(\hat{\lambda}_{r,0},\ldots,\hat{\lambda}_{r,J+1})) \quad (r=1,\ldots,R)$$

where n_r is the number of sample units having the rth covariate level and

$$\hat{\lambda}_{r,j} = \phi_j\cdot\frac{1}{1+e^{x_{i_r}\hat{\beta}}} + \phi_{j-1}\cdot\frac{e^{x_{i_r}\hat{\beta}}}{1+e^{x_{i_r}\hat{\beta}}}.$$

Then, $\hat{\beta}_{(b)}$ is the MLE corresponding to the new frequencies $n_1^{(b)}(r,j)$ and can be computed as above.

We now switch to stochastic covariates. Here, each person in the sample is first requested to reveal his or her outcomes of the nonsensitive covariates, and second to give a reply $S=Z+Y$ according to the ICT without control group. The case of stochastic explanatory variables is likely more relevant for practice than the case of deterministic covariates. However, we have initially presented the deterministic case, because this case is mathematically simpler and the estimation for stochastic exogenous characteristics can be traced back to this case. Let the random variable X_{ij} be the ith sample unit's outcome of the jth covariate and define the random vector $X_i=(X_{i1},\ldots,X_{ip})$. We have to incorporate the stochastic character of the covariates into our assumptions. In particular, (D1)–(D4) change to

(S1) $(Y_1, X_1), \ldots, (Y_n, X_n)$ are iid.
(S2) $Z_1, \ldots, Z_n$ are iid with $Z_i \sim Z$.
(S3) The two quantities $\begin{pmatrix} Y_1, & X_1 \\ \vdots & \vdots \\ Y_n, & X_n \end{pmatrix}$ and $\begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix}$ are independent.
(S4) There is a $\beta \in \mathbb{R}^p$ with $\mathbb{P}(Y_i = 1 | X_i = x_i) = \frac{e^{x_i\beta}}{1+e^{x_i\beta}}$ $(i = 1, \ldots, n)$.

(S1)–(S3) are satisfied if (Y, X) and Z are independent and the respondents are drawn by SRSWR from the universe. The observed data log-likelihood is given by (a constant is ignored)

$$\begin{aligned} l_{obs}(\beta) &= \sum_{i=1}^{n} \log \mathbb{P}(S_i = s_i | X_i = x_i) \\ &= \sum_{i=1}^{n} \log \left[\phi_{s_i} \cdot \frac{1}{1 + e^{x_i\beta}} + \phi_{s_i - 1} \cdot \frac{e^{x_i\beta}}{1 + e^{x_i\beta}} \right]. \end{aligned} \tag{10}$$

A comparison with (9) makes clear that the maximum of (10) can be obtained by maximizing an observed data log-likelihood corresponding to certain data according to the case of deterministic covariates. How this can be done, is explained above. Estimated standard errors for the components of $\hat{\beta}$ given the observed covariate levels can be calculated by a bootstrap procedure analog to the case of deterministic exogenous variables.

3.7. Numerical example for covariate extension

In this subsection, we illustrate the proposed methods from Section 3.6 in a simple, artificial example. We consider $Y \in \{0, 1\}$ where we may have $Y = 1$ if a person conducts undeclared work and $Y = 0$ else. Let us suppose that we have $p = 4$ covariates. The first covariate is a constant equal to 1 while the other covariates each can attain the values 0 and 1. The covariates 2–4 could have the following meanings. The second covariate may equal 1 if the person has a regular occupation of at least 30 hours per week. The third covariate may describe the gender (0: female/1: male). The fourth covariate may attain value 1 if the person expects a great chance of being caught when conducting undeclared work. Then, the covariate vector $(1, 1, 0, 1)$ represents a woman having a main occupation of 30 or more hours per week and evaluating the probability of being caught when conducting moonlighting to be large. Say, the number Z of affirmative responses to the nonsensitive questions is binomially distributed with parameters 4 and 0.5. Let us now assume that we have made the observations as given in the left part of Table 4.

Under the assumptions (S1)–(S4), we obtain the MLE

$$\hat{\beta} = \begin{pmatrix} 1.2270 & -2.3233 & 1.6568 & -3.3706 \end{pmatrix}^\top$$

for β with 87 iteration of the EM algorithm. With this $\hat{\beta}$, we estimate the probability that a person conducts undeclared work given the covariates with

Table 4
Observed distribution of the scrambled answers S and estimated probabilities of the sensitive variable Y for each covariate level

covariate level X	observed frequency $S=0$	$S=1$	$S=2$	$S=3$	$S=4$	$S=5$	estimated probability $Y=0$	$Y=1$
$(1,0,0,0)$	1	8	21	26	15	3	0.2267	0.7733
$(1,1,0,0)$	9	37	64	53	21	3	0.7496	0.2504
$(1,0,1,0)$	0	6	19	27	18	4	0.0530	0.9470
$(1,0,0,1)$	1	6	9	7	2	0	0.8951	0.1049
$(1,1,1,0)$	1	5	11	12	7	1	0.3634	0.6366
$(1,1,0,1)$	4	15	23	16	4	0	0.9886	0.0114
$(1,0,1,1)$	1	4	8	7	3	1	0.6193	0.3807
$(1,1,1,1)$	1	3	5	3	1	0	0.9432	0.0568

the help of the relation (S4). These estimated probabilities are given in the right part of Table 4. For our contrived data and the corresponding estimation, the following interpretations hold: Men are more involved in undeclared work than women, because $\hat{\beta}_3$ is positive. People expecting a large chance of being caught are less often involved in moonlighting than persons who expect a small chance of being caught, because $\hat{\beta}_4$ is negative. The largest probability for committing this kind of social fraud appears for men without main occupation of at least 30 hours per week who assess the chance of being caught to be small. For women with regular occupation of at least 30 hours per week who expect a large chance of being caught, the probability for undeclared work is smallest.

4. Extension of the ICT to polychotomous sensitive attributes

Sometimes an investigator may be interested in a sensitive characteristic with more than two categories. Examples for such variables are income (divided in classes) and the number of times a person has evaded taxes. In this section, let Y be a sensitive attribute with an arbitrary number k of categories coded with $0,\dots,k-1$. As before U_j $(j=1,\dots,J)$ stands for an innocuous attribute, but we now allow that U_j attains values $0,\dots,k_j$ $(k_j \geq 1)$. For instance, we may define $U_1 \in \{0,1,2\}$ where $U_1=0$ if a person did not visit a foreign country last year, $U_1=1$ if a person visited a foreign country once last year, and $U_1=2$ if a person visited a foreign country two or more times last year. Analog to Sections 2 and 3, we define Z to be the sum of the nonsensitive variables, i.e., $Z=U_1+\cdots+U_J$. In this section, we first investigate the case in which the distribution of Z is not known and establish an item count technique with control and treatment group that enables the estimation of the probability masses of the multicategorical Y. A Bayesian extension and regression analysis are also described. Subsequently, we consider a known distribution of Z. For this case, we show that a control group is not necessary and derive the corresponding estimates.

4.1. Method and estimation

We divide the interviewees in two groups, one control group and one treatment group where each respondent has a chance of 50% to be assigned to the

TABLE 5

Example of a questionnaire for the polychotomous ICT that is shown to the respondents in the treatment group. The questionnaire should be accompanied with an instruction like "For each question, please think about your answer category. Subsequently, compute the sum of the answer categories that apply to you. Report this sum and nothing else." By deleting the question on tax evasion, we obtain the questionnaire for persons in the control group

Question	Answer category	Answer statement
Did you attend a religious service last month?	**0**	no
	1	once
	2	twice or more
Did you visit a foreign country last year?	**0**	no
	1	yes
Is your telephone number even?	**0**	no
	1	yes
What is your favourite sport?	**0**	soccer
	1	handball
	2	athletics
	3	swimming
	4	other
Have you evaded taxes last year?	**0**	I don't have evaded taxes.
	1	I have evaded taxes, but not more than 1000 Euro.
	2	I have evaded taxes in an amount of more than 1000 Euro.

treatment group. Each respondent in the control group is supplied with a questionnaire with the J nonsensitive questions. The interviewee is introduced to get the outcome for each question straight in his or her mind and reveal only the sum of the outcomes. In the treatment group, every respondent receives the same nonsensitive questions and additionally a question concerning the critical Y. In this group, the demanded answer is the overall sum of outcomes for the nonsensitive variables and the sensitive variable. An example of a questionnaire for the treatment group is provided in Table 5. Such a table should be accompanied with the clear instruction that only a total has to be reported. Additionally, an example of an answer may be helpful.

Contrary to Section 3, we need a treatment indicator for the current setup. Let $T \in \{0, 1\}$ be this indicator. Then, requested answer is $S = \sum_{j=1}^{J} U_j + T \cdot Y = Z + T \cdot Y$. Say $Z \in \{0, \dots, k_Z - 1\}$ where $k_Z - 1 = k_1 + \cdots + k_J$. Consequently, $S \in \{0, \dots, k_Z + k - 2\}$, that is, there are $k_Z + k - 1$ answer categories where the replies $k_Z, \dots, k_Z + k - 2$ can only emerge in the treatment group. It must be mentioned that the ceiling effect explained in Section 2 propagates itself to the polychotomous case. In particular, for a respondent in the treatment group, the answers $k_Z, \dots, k_Z + k - 2$ restrict the possible Y-values. E.g., for $k = 3$ and $k_Z = 7$, we have that $S = 8$ implies $Y = 2$ while $S = 7$ implies $Y \geq 1$. To reduce the ceiling effect, it is appealing to select control items where one can expect only few persons possessing values of Z greater or equal than $k_Z - k + 1$.

Adapting the notation from Sections 2 and 3, we now have $\pi = (\pi_0, \ldots, \pi_{k-1})^\top$ and $\phi = (\phi_0, \ldots \phi_{k_Z-1})^\top$. We again assume independence of Z and Y and consider SRSWR of size n. We define $T_i = 1$ if the ith sample unit belongs to the treatment group and $T_i = 0$ else. The T_i are collected in $\mathbf{T} = (T_1, \ldots, T_n)$. Furthermore, let t_i denote the realization of T_i and set $\mathbf{t} = (t_1, \ldots, t_n)$. The observed data log-likelihood in our polychotomous case is given by

$$l_{obs}(\pi, \phi; \mathbf{s}, \mathbf{t}) = \sum_{i=1}^{n} \log \mathbb{P}(S_i = s_i | T_i = t_i) = \sum_{i=1}^{n} \log \left[\sum_{j=0}^{k-1} \phi_{s_i - t_i j} \cdot \pi_j \right]. \quad (11)$$

Here, the additive constant $\sum_{i=1}^{n} \log P(T_i = t_i) = n \cdot \log 0.5$ is ignored, since it is irrelevant for the maximization. At other places in this paper, similar ignorings occur although this may be not explicitly emphasized. The complete data log-likelihood equals

$$l_{com}(\pi, \phi) = l_{com}(\pi, \phi; \mathbf{y}, \mathbf{z}, \mathbf{s}, \mathbf{t}) = \sum_{i=1}^{n} \log \mathbb{P}(Y_i = y_i) + \sum_{i=1}^{n} \log \mathbb{P}(Z_i = z_i)$$
$$= \sum_{j=0}^{k-1} \log \pi_j \cdot \sum_{i=1}^{n} 1_{\{j\}}(y_i) + \sum_{j=0}^{k_Z-1} \log \phi_j \cdot \sum_{i=1}^{n} 1_{\{j\}}(z_i).$$

The maximization of l_{obs} can again be conducted with the EM algorithm. In the E step of iteration $t+1$, we estimate the complete data log-likelihood by

$$\begin{aligned}
\widehat{l_{com}}(\pi, \phi) &= \mathbb{E}_t(l_{com}(\pi, \phi; \mathbf{Y}, \mathbf{Z}, \mathbf{S}, \mathbf{T}) \,|\, \mathbf{S} = \mathbf{s}, \mathbf{T} = \mathbf{t}) \\
&= \sum_{j=0}^{k-1} \log \pi_j \cdot \sum_{i=1}^{n} \mathbb{E}_t(1_{\{j\}}(Y_i) | \mathbf{S} = \mathbf{s}, \mathbf{T} = \mathbf{t}) \\
&+ \sum_{j=0}^{k_Z-1} \log \phi_j \cdot \sum_{i=1}^{n} \mathbb{E}_t(1_{\{j\}} Z_i | \mathbf{S} = \mathbf{s}, \mathbf{T} = \mathbf{t}) \\
&=: \sum_{j=0}^{k-1} \log \pi_j \cdot v_j^{(t)} + \sum_{j=0}^{k_Z-1} \log \phi_j \cdot w_j^{(t)}.
\end{aligned}$$

Here, for $j = 0, \ldots, k-1$ respectively $j = 0, \ldots, k_Z - 1$, the identities

$$\mathbb{E}_t(1_{\{j\}}(Y_i) | \mathbf{S} = \mathbf{s}, \mathbf{T} = \mathbf{t}) \quad = \quad \mathbb{P}_t(Y_i = j | S_i = s_i, T_i = t_i) = \frac{\phi_{s_i - t_i j}^{(t)} \cdot \pi_j^{(t)}}{\sum_{l=0}^{k-1} \phi_{s_i - t_i l}^{(t)} \cdot \pi_l^{(t)}}$$

and

$$\mathbb{E}_t(1_{\{j\}}(Z_i) | \mathbf{S} = \mathbf{s}, \mathbf{T} = \mathbf{t}) \quad = \quad \sum_{l=0}^{k-1} 1_{\{s_i - t_i l\}}(j) \cdot \mathbb{P}_t(Y_i = l | S_i = s_i, T_i = t_i)$$

hold. In the M step of iteration $t+1$, we obtain $\pi^{(t+1)}$ and $\phi^{(t+1)}$ by maximizing $\widehat{l_{com}}$. Here, we have

$$\pi_j^{(t+1)} = \frac{v_j^{(t)}}{v_0^{(t)} + \cdots + v_{k-1}^{(t)}}, \quad \phi_j^{(t+1)} = \frac{w_j^{(t)}}{w_0^{(t)} + \cdots + w_{k_Z-1}^{(t)}}.$$

This algorithm can be programmed conveniently. For this purpose, we point out that

$$\begin{aligned} \begin{pmatrix} v_0^{(t)} \\ \vdots \\ v_{k-1}^{(t)} \end{pmatrix} &= n_C \cdot \pi^{(t)} + \left(\boldsymbol{\phi}^{(t)} \,.^* \left[\begin{pmatrix} 1/\lambda_0^{(t)} \\ \vdots \\ 1/\lambda_{k_Z+k-2}^{(t)} \end{pmatrix} \cdot (\pi_0^{(t)}, \ldots, \pi_{k-1}^{(t)}) \right] \right)^\top \cdot n_1^\top \\ &=: n_C \cdot \pi^{(t)} + P_1^{(t)} \cdot n_1^\top \end{aligned} \tag{12}$$

where n_C is the size of the control group, $\boldsymbol{\phi}^{(t)}$ is a $(k_Z + k - 1) \times k$ matrix whose entry (i, j) is $\phi_{i-j}^{(t)}$ for $i = 0, \ldots, k_Z + k - 2$; $j = 0, \ldots, k - 1$, and $\lambda^{(t)} = (\lambda_0^{(t)}, \ldots, \lambda_{k_Z+k-2}^{(t)})^\top = \boldsymbol{\phi}^{(t)} \cdot (\pi_0^{(t)}, \ldots, \pi_{k-1}^{(t)})^\top$. Moreover, n_{1i} represents the absolute frequency of answer i among the respondents in the treatment group and $n_1 = (n_{10}, \ldots, n_{1,k_Z+k-2})$. Regarding the $w_j^{(t)}$, we introduce the $k_Z \times (k_Z + k - 1)$ matrix $\tilde{P}^{(t)}$ whose component (i, j) is equal to entry $(j - i, j)$ of the matrix $P_1^{(t)}$ for $i = 0, \ldots, k_Z - 1$ and $j = 0, \ldots, k_Z + k - 2$. Then, it follows

$$(w_0^{(t)}, \ldots, w_{k_Z-1}^{(t)})^\top = n_0^\top + \tilde{P}^{(t)} \cdot n_1^\top \tag{13}$$

with $n_0 = (n_{00}, \ldots, n_{0,k_Z-1})$ describing the observed answer distribution in the control group, that is, n_0 is the analog of n_1 for the control group. As initial values, we may employ $\pi^{(0)}$ and $\phi^{(0)}$ that each consist of identical entries. The algorithm stops if the deviation between $(\pi^{(t)}, \phi^{(t)})$ and the successor $(\pi^{(t+1)}, \phi^{(t+1)})$ is sufficiently small. The generated sequence $(\pi^{(t)}, \phi^{(t)})_{t \in \mathbb{N}_0}$ yields a nondecreasing sequence $(l_{obs}(\pi^{(t)}, \phi^{(t)}; \mathbf{s}, \mathbf{t}))_{t \in \mathbb{N}_0}$. The last M step delivers the estimate $\hat{\theta} = (\hat{\pi}^\top, \hat{\phi}^\top)^\top$.

Estimated standard errors of the components of $\hat{\theta}$ and confidence intervals for components of $\theta = (\pi^\top, \phi^\top)^\top$, can be derived similar to Section 3 from B bootstrap replications of $\hat{\theta}$. The bth reproduction is generated by simulating the size of the control group $n_C^{(b)} \sim Bin(n, 0.5)$, drawing new frequencies of the answers in the groups by $n_0^{(b)} \sim Multinomial(n_C^{(b)}, \hat{\phi}^\top)$ and $n_1^{(b)} \sim Multinomial(n - n_C^{(b)}, (\hat{\lambda}_0, \ldots, \hat{\lambda}_{k_Z+k-2}))$ with $\hat{\lambda}_i = \sum_{j=0}^{k-1} \hat{\phi}_{i-j} \cdot \hat{\pi}_j$. Then, $\hat{\theta}^{(b)}$ is the MLE corresponding to $n_0^{(b)}$ and $n_1^{(b)}$.

4.2. Numerical illustration

This subsection provides a numerical example for the polychotomous ICT. The following setup and data are artificial, but suffice to illustrate the estimation methods according to Subsection 4.1. Let us assume that a questionnaire as in

Table 6
Observed distribution of the answers S

group / answer S	0	1	2	3	4	5	6	7	8	9	10
control	30	71	90	92	81	60	42	26	8	–	–
treatment	21	56	80	90	84	68	49	32	15	4	1

Table 5 is used in the treatment group while this questionnaire without the last (i.e., the sensitive) question is used in the control group. Consequently, we have $J = 4$ nonsensitive control items, Z can attain the $k_Z = 9$ values $0, \dots, 8$, and the sensitive attribute on tax evasion possesses the three categories 0, 1, and 2. Thus, in the treatment group, $k_Z + k - 1 = 11$ different answers are possible (namely, the answers $0, \dots, 10$). Furthermore, let us assume that the sample size equals $n = 1000$, where each 500 interviewees are assigned to the control and treatment group, and that we have observed the frequencies of the answers as in Table 6.

Then, the EM algorithm delivers the MLEs $\hat{\pi} = (0.7148, 0.1761, 0.1091)^\top$ and

$$\hat{\phi} = (0.0595, 0.1420, 0.1800, 0.1847, 0.1621, 0.1205, 0.0836, 0.0514, 0.0162)^\top.$$

According to $\hat{\pi}$, the majority of inhabitants would pay the taxes correctly while an estimated proportion of 11% would have evaded more than 1000 Euro taxes. Moreover, we obtained bootstrap standard errors of the components $SE(\hat{\pi}) = (0.1173, 0.1641, 0.0786)^\top$ and e.g. a 90% bootstrap confidence interval for π_1 given by [0.4939, 0.8651].

4.3. Bayes extension

We establish Bayesian estimates for the polychotomous ICT from Section 4.1 in this subsection. Here, we modify the considerations from Subsection 3.5. The true π and ϕ are treated as realizations of random quantities $(\Pi_0, \dots, \Pi_{k-1})^\top$ and $(\Phi_0, \dots, \Phi_{k_Z-1})^\top$, respectively. As prior density for $(\Pi_0, \dots, \Pi_{k-2})$, we set

$$f_{\Pi_0,\dots,\Pi_{k-2}}(\pi_0, \dots, \pi_{k-2}) = constant \cdot \pi_0^{\delta_0 - 1} \cdots \pi_{k-1}^{\delta_{k-1}-1}$$

for $\pi_0, \dots, \pi_{k-2} \in [0, 1]$, $\pi_0 + \dots + \pi_{k-2} \leq 1$, $\pi_{k-1} = 1 - \pi_0 - \dots - \pi_{k-2}$ and $\delta_i > 0$, that is, we have a Dirichlet distribution with parameters $\delta_0, \dots, \delta_{k-1}$. The Dirichlet distribution is a multivariate extension of the Beta distribution. We also apply the Dirichlet distribution for the prior of $(\Phi_0, \dots, \Phi_{k_Z-2})$, more precisely, we assume $(\Phi_0, \dots, \Phi_{k_Z-2})$ to have a Dirichlet distribution with parameters $\varepsilon_0, \dots, \varepsilon_{k_Z-1}$. As overall prior density, we use

$$\begin{aligned} &f_{\Pi_0,\dots,\Pi_{k-2},\Phi_0,\dots,\Phi_{k_Z-2}}(\pi_0, \dots, \pi_{k-2}, \phi_0, \dots, \phi_{k_Z-2}) \\ &\qquad = f_{\Pi_0,\dots,\Pi_{k-2}}(\pi_0, \dots, \pi_{k-2}) \cdot f_{\Phi_0,\dots,\Phi_{k_Z-2}}(\phi_0, \dots, \phi_{k_Z-2}). \end{aligned}$$

The advantages of this prior are similar to those of the prior in Subsection 3.5. In particular, the prior contains information equivalent to $\delta_0 + \dots + \delta_{k-1} - k$

observations on Y where $Y = i$ occurs $\delta_i - 1$ times and $\varepsilon_0 + \cdots + \varepsilon_{k_Z-1} - k_Z$ additional data on Z among which $Z = i$ appears $\varepsilon_i - 1$ times. Moreover, a researcher's guesses for π and ϕ can be converted into a concrete prior where the certainty is reflected and the procedures of EM and data augmentation algorithm are relatively simple. For the density of the complete data given the parameter, we define

$$f_{\mathbf{Y},\mathbf{S},\mathbf{T}\,|\,\Pi_0,\ldots,\Pi_{k-2},\Phi_0,\ldots,\Phi_{k_Z-2}}(\mathbf{y},\mathbf{s},\mathbf{t}\,|\,\pi_0,\ldots,\pi_{k-2},\phi_0,\ldots,\phi_{k_Z-2}) = \prod_{i=1}^{n} \pi_{y_i} \cdot \phi_{s_i - t_i y_i} \cdot \frac{1}{2}$$

where, of course, $\pi_{k-1} = 1 - \pi_0 - \cdots - \pi_{k-2}$ and $\phi_{k_Z-1} = 1 - \phi_0 - \cdots - \phi_{k_Z-2}$ hold. Regarding the calculation of the mode of the density

$$f_{\Pi_0,\ldots,\Pi_{k-2},\Phi_0,\ldots,\Phi_{k_Z-2}\,|\,\mathbf{S},\mathbf{T}}(\pi_0,\ldots,\pi_{k-2},\phi_0,\ldots,\phi_{k_Z-2}\,|\,\mathbf{s},\mathbf{t})$$

via EM algorithm, function (8), which corresponds to the E step, changes to

$$\begin{aligned}(\pi_0,\ldots,\pi_{k-2},\phi_0,\ldots,\phi_{k_Z-2}) \mapsto &\sum_{j=0}^{k-1} \log \pi_j \cdot v_j^{(t)} + \sum_{j=0}^{k_Z-1} \log \phi_j \cdot w_j^{(t)} \\ &+ \sum_{j=0}^{k-1} \log \pi_j \cdot (\delta_j - 1) + \sum_{j=0}^{k_Z-1} \log \phi_j \cdot (\varepsilon_j - 1)\end{aligned} \tag{14}$$

where the $v_j^{(t)}$ are from (12) and the $w_j^{(t)}$ are from (13). Obviously, (14) comprises a part that corresponds to the estimated complete data log-likelihood for the non-Bayes case and a part that belongs to the prior.

The DA algorithm leads to realizations $(\mathbf{y}^{(t)}, \pi_0^{(t)}, \ldots, \pi_{k-2}^{(t)}, \phi_0^{(t)}, \ldots, \phi_{k_Z-2}^{(t)})_{t\geq 1}$ of a Markov chain $(\mathbf{Y}^{(t)}, \Pi_0^{(t)}, \ldots, \Pi_{k-2}^{(t)}, \Phi_0^{(t)}, \ldots, \Phi_{k_Z-2}^{(t)})_{t\geq 1}$ that converges in distribution to $(\mathbf{Y}, \Pi_0, \ldots, \Pi_{k-2}, \Phi_0, \ldots, \Phi_{k_Z-2})$ given $\mathbf{s}$ and $\mathbf{t}$. In the I step of iteration $t+1$, we generate the vector $\mathbf{y}^{(t+1)}$ from

$$f_{\mathbf{Y}\,|\,\Pi_0,\ldots,\Pi_{k-2},\Phi_0,\ldots,\Phi_{k_Z-2},\mathbf{S},\mathbf{T}}(\mathbf{y}\,|\,\pi_0^{(t)},\ldots,\pi_{k-2}^{(t)},\phi_0^{(t)},\ldots,\phi_{k_Z-2}^{(t)},\mathbf{s},\mathbf{t}) = \prod_{i=1}^{n} \frac{\pi_{y_i}^{(t)} \cdot \phi_{s_i-t_i y_i}^{(t)}}{\sum_{j=0}^{k-1} \pi_j^{(t)} \cdot \phi_{s_i - t_i j}^{(t)}}.$$

In the subsequent posterior step (P step) of iteration $t+1$, we draw new parameters from the distribution of $(\Pi_0, \ldots, \Pi_{k-2}, \Phi_0, \ldots, \Phi_{k_Z-2})$ given $\mathbf{Y} = \mathbf{y}^{(t+1)}$, $\mathbf{S} = \mathbf{s}$, $\mathbf{T} = \mathbf{t}$. The density of this distribution is the product of the density corresponding to a Dirichlet distribution with parameters $m_0^{(t+1)} + \delta_0, \ldots, m_{k-1}^{(t+1)} + \delta_{k-1}$ and the density of a Dirichlet distribution with parameters $q_0^{(t+1)} + \varepsilon_0, \ldots, q_{k_Z-1}^{(t+1)} + \varepsilon_{k_Z-1}$. Here, $m_j^{(t+1)}$ denotes the number how often the value j appears among $y_1^{(t+1)}, \ldots, y_n^{(t+1)}$ and $q_j^{(t+1)}$ represents the number how often outcome j occurs among $z_1^{(t+1)}, \ldots, z_n^{(t+1)}$ where we set $z_i^{(t+1)} = s_i - t_i y_i^{(t+1)}$.

With the help of the generated realizations of the Markov chain, we are able to simulate expectations and quantiles of the distribution of Π_i given $\mathbf{s}$ and $\mathbf{t}$ as well as of P_i given $\mathbf{s}$ and $\mathbf{t}$ where $P_i = n^{-1}\sum_{j=1}^{n} 1_{\{Y_j=i\}}$. These simulations proceed analog to Subsection 3.5, in which the ICT for binary target variables without control group is under study. We can also simulate expectations and quantiles of the distribution of Φ_i given $\mathbf{s}$ and $\mathbf{t}$. However, these quantities are typically of lower interest, because we are mainly interested in the sensitive variable.

4.4. Regression analysis

Regarding the ICT for binary sensitive variables according to Miller (1984, [14]), methods for regression analysis are proposed e.g. in Imai (2011, [12]) and in Kuha and Jackson (2014, [13]). An important element of these methods is that a structure model for the control items has to be specified. In this subsection, we extend the available literature by techniques that enable the investigation of the influence of nonsensitive covariates on a multicategorical sensitive item on which data are collected by the ICT from Subsection 4.1. Let us first consider deterministic covariates. In this case, values of the covariates are determined by the researcher and persons having these values are searched. Each person is randomly assigned either to the control or to the treatment group and is requested to give an answer according to the polychotomous ICT from 4.1. Let $x_i \in \mathbb{R}^{1\times p}$ be a vector whose jth entry represents the ith interviewee's value of covariate j ($i = 1,\dots,n$; $j = 1,\dots,p$). We assume:

- (D1') The n vectors $(Y_1, Z_1, T_1),\dots,(Y_n, Z_n, T_n)$ are independent.
- (D2') For all $i = 1,\dots,n$, we have: T_i and (Y_i, Z_i) are independent and $\mathbb{P}(T_i = 1) = 1/2$.
- (D3') There is a $\beta = (\beta^{(1)^\top},\dots,\beta^{(k-1)^\top})^\top$ with $\beta^{(j)} \in \mathbb{R}^{p\times 1}$ and

$$\mathbb{P}(Y_i = j) = \frac{e^{x_i\beta^{(j)}}}{1 + e^{x_i\beta^{(1)}} + \cdots + e^{x_i\beta^{(k-1)}}} \quad (j = 1,\dots,k-1).$$

Hence, we are in the situation of a multivariate logistic regression model for the influence of the exogenous quantities on the stigmatizing variable. The assumptions (D1') and (D2') are fulfilled in the following case: for any covariate level, the respondents are drawn by SRSWR out of the population units having this covariate level such that the selection is conducted independent of the selection for other covariate levels; the drawn persons are randomly assigned to a group e.g. by flipping a fair coin. Notice, we allow dependence between Y_i and Z_i in this subsection. For $s_i \in \{0,\dots,k_Z + k - 2\}$ and $t_i \in \{0, 1\}$, it is true that

$$\begin{aligned}\log \mathbb{P}\left(\bigcap_{i=1}^{n}\{S_i = s_i, T_i = t_i\}\right) \\ = \sum_{i=1}^{n}\left(\log\left[\sum_{j=0}^{k-1}\mathbb{P}(Z_i = s_i - t_i j \,|\, Y_i = j)\cdot\mathbb{P}(Y_i = j)\right] + \log 0.5\right).\end{aligned}$$

As in Kuha and Jackson (2014, [13]) for the binary ICT, we make use of a model for the probabilities $\mathbb{P}(Z_i = z \mid Y_i = y)$. In particular, we consider a multinomial logistic regression set up (compare Appendix A3 in Kuha and Jackson (2014, [13]) for the binary ICT). Other modelings lead to similar estimation steps. Formally, we assume that a $\psi = (\psi^{(1)^\top}, \dots, \psi^{(k_Z-1)^\top})^\top$ exists such that

$$\mathbb{P}(Z_i = z \mid Y_i = y) = \frac{e^{v(x_i,y)\cdot\psi^{(z)}}}{1 + e^{v(x_i,y)\cdot\psi^{(1)}} + \cdots + e^{v(x_i,y)\cdot\psi^{(k_Z-1)}}} \tag{15}$$

holds for $z = 1, \dots, k_Z - 1$. In this equation, v is a map specified by the researcher and the range of v determines the length of the row vector $\psi^{(l)}$ ($l = 1, \dots, k_Z - 1$). We give some examples for v. For $v(x, y) = (x, y) \in \mathbb{R}^{1\times(p+1)}$, the distribution of Z_i depends on the nonsensitive covariates and the sensitive item. Some authors recommend that the control items should not be totally unrelated to the sensitive item (e.g., Chaudhuri and Christofides (2007, Section 3, [4])). For such a situation a v with $v(x, y)$ depending on y is helpful. In the case, $v(x, y) = x$, Z_i is independent of the sensitive characteristic and for $v(x, y) = 1$, Z_i is independent of both the innocent covariates and the stigmatizing variable. The observed data log-likelihood is given by

$$l_{obs}(\beta, \psi) = \sum_{i=1}^{n} \log \left[\sum_{j=0}^{k-1} \mathbb{P}(Z_i = s_i - t_i j \mid Y_i = j) \cdot \mathbb{P}(Y_i = j) \right] \tag{16}$$

whereas the complete data log-likelihood has the form

$$\begin{aligned} l_{com}(\beta, \psi) &= \sum_{i=1}^{n} \log \mathbb{P}(Y_i = y_i) + \sum_{i=1}^{n} \log \mathbb{P}(Z_i = z_i \mid Y_i = y_i) \\ &=: l_1(\beta) + l_2(\psi). \end{aligned} \tag{17}$$

Again, the EM algorithm is beneficial to maximize (16). Let estimates $\beta_{(t)} = (\beta_{(t)}^{(1)^\top}, \dots, \beta_{(t)}^{(k-1)^\top})^\top$ and $\psi_{(t)} = (\psi_{(t)}^{(1)^\top}, \dots, \psi_{(t)}^{(k_Z-1)^\top})^\top$ for β and ψ be available from iteration t. In the expectation step of iteration $t+1$, $l_1(\beta)$ and $l_2(\psi)$ are replaced by certain conditional expectations $l_1^{(t)}(\beta)$ and $l_2^{(t)}(\psi)$. In detail, we have with $\beta^{(0)}$ being a vector of zeros

$$\begin{aligned} l_1^{(t)}(\beta) = &\sum_{i=1}^{n_C} \sum_{j=0}^{k-1} \mathbb{P}_t(Y_i = j \mid S_i = s_i, T_i = 0) \cdot \log \frac{e^{x_i\beta^{(j)}}}{1 + e^{x_i\beta^{(1)}} + \cdots + e^{x_i\beta^{(k-1)}}} \\ &+ \sum_{i=n_C+1}^{n} \sum_{j=0}^{k-1} \mathbb{P}_t(Y_i = j \mid S_i = s_i, T_i = 1) \cdot \log \frac{e^{x_i\beta^{(j)}}}{1 + e^{x_i\beta^{(1)}} + \cdots + e^{x_i\beta^{(k-1)}}} \\ &=: l_{10}^{(t)}(\beta) + l_{11}^{(t)}(\beta) \end{aligned}$$

where we assume without loss of generality that the sample units $i = 1, \dots, n_C$ are assigned to the control group while the units $i = n_C + 1, \dots, n$ belong to the treatment group. That is, $l_{10}^{(t)}(\beta)$ and $l_{11}^{(t)}(\beta)$ correspond to the control and treatment group, respectively. Further, $l_{10}^{(t)}(\beta) =$

$$\sum_{j=0}^{k-1}\sum_{r=1}^{R_0}\sum_{s=0}^{k_Z-1} n_0(r,s)\cdot \mathbb{P}_t(Y_{i_{0r}}=j\,|\,S_{i_{0r}}=s,T_{i_{0r}}=0)\cdot \log\frac{e^{x_{i_{0r}}\beta^{(j)}}}{1+\cdots+e^{x_{i_{0r}}\beta^{(k-1)}}} \quad (18)$$

holds. Here, we assume that we have $R_0 \leq n_C$ covariate levels for respondents in the control group and that sample unit $i_{0r} \in \{1,\ldots,n_C\}$ possesses the rth covariate level. Moreover, we denote the number how often answer s occurs among the respondents in the control group with covariate level r by $n_0(r,s)$. Concerning (18), we have

$$\mathbb{P}_t(Y_{i_{0r}}=j\,|\,S_{i_{0r}}=s,T_{i_{0r}}=0)=\frac{\mathbb{P}_t(Z_{i_{0r}}=s\,|\,Y_{i_{0r}}=j)\cdot\mathbb{P}_t(Y_{i_{0r}}=j)}{\sum_{l=0}^{k-1}\mathbb{P}_t(Z_{i_{0r}}=s\,|\,Y_{i_{0r}}=l)\cdot\mathbb{P}_t(Y_{i_{0r}}=l)} \text{ with}$$

$$\mathbb{P}_t(Z_{i_{0r}}=s\,|\,Y_{i_{0r}}=l)=\frac{e^{v(x_{i_{0r}},l)\cdot\psi_{(t)}^{(s)}}}{1+e^{v(x_{i_{0r}},l)\cdot\psi_{(t)}^{(1)}}+\cdots+e^{v(x_{i_{0r}},l)\cdot\psi_{(t)}^{(k_Z-1)}}} \text{ and}$$

$$\mathbb{P}_t(Y_{i_{0r}}=l)=\frac{e^{x_{i_{0r}}\beta_{(t)}^{(l)}}}{1+e^{x_{i_{0r}}\beta_{(t)}^{(1)}}+\cdots+e^{x_{i_{0r}}\beta_{(t)}^{(k-1)}}}.$$

For these identities, we define $\psi_{(t)}^{(0)}$ and $\beta_{(t)}^{(0)}$ to be vectors consisting only of zeros. For the function $l_{11}^{(t)}$, it is true that $l_{11}^{(t)}(\beta) =$

$$\sum_{j=0}^{k-1}\sum_{r=1}^{R_1}\sum_{s=0}^{k_Z+k-2} n_1(r,s)\cdot\mathbb{P}_t(Y_{i_{1r}}=j\,|\,S_{i_{1r}}=s,T_{i_{1r}}=1)\cdot\log\frac{e^{x_{i_{1r}}\beta^{(j)}}}{1+\cdots+e^{x_{i_{1r}}\beta^{(k-1)}}}$$

where

$$\mathbb{P}_t(Y_{i_{1r}}=j\,|\,S_{i_{1r}}=s,T_{i_{1r}}=1)=\frac{\mathbb{P}_t(Z_{i_{1r}}=s-j\,|\,Y_{i_{1r}}=j)\cdot\mathbb{P}_t(Y_{i_{1r}}=j)}{\sum_{l=0}^{k-1}\mathbb{P}_t(Z_{i_{1r}}=s-l\,|\,Y_{i_{1r}}=l)\cdot\mathbb{P}_t(Y_{i_{1r}}=l)}$$

and the probabilities contained in this fraction come from (15) and (D3') by working with $\psi_{(t)}$ and $\beta_{(t)}$ instead of ψ and β. Additionally, R_1 denotes the number of covariate levels in the treatment group, sample unit $i_{1r} \in \{n_C+1,\ldots,n\}$ is a person having the rth covariate level, and $n_1(r,s)$ is the absolute frequency of interviewees in the treatment group with covariate level r giving answer s. Let us now consider $l_2(\psi)$. Partitioning the respondents in control and treatment group yields $l_2(\psi) =$

$$\sum_{i=1}^{n_C}\log\mathbb{P}(Z_i=z_i\,|\,Y_i=y_i)+\sum_{i=n_C+1}^{n}\log\mathbb{P}(Z_i=z_i\,|\,Y_i=y_i)=:l_{20}(\psi)+l_{21}(\psi).$$

The first summand can be written as

$$l_{20}(\psi)=\sum_{i=1}^{n_C}\sum_{j=0}^{k-1}\sum_{s=0}^{k_Z-1}1_{\{j\}}(y_i)\cdot 1_{\{s\}}(s_i)\cdot\log\mathbb{P}(Z_i=s\,|\,Y_i=j)$$

while the second summand is equal to

$$l_{21}(\psi) = \sum_{i=n_C+1}^{n} \sum_{j=0}^{k-1} \sum_{s=0}^{k_Z-1} 1_{\{j\}}(y_i) \cdot 1_{\{s+j\}}(s_i) \cdot \log \mathbb{P}(Z_i = s \mid Y_i = j).$$

In the E step of iteration $t+1$, we substitute $l_{20}(\psi)$ and $l_{21}(\psi)$ by their conditional expectations given the observed data and calculated under the parameters from iteration t and obtain

$$\begin{aligned} l_{20}^{(t)}(\psi) = \sum_{j=0}^{k-1} \sum_{r=1}^{R_0} \sum_{s=0}^{k_Z-1} & n_0(r,s) \cdot \mathbb{P}_t(Y_{i_{0r}} = j \mid S_{i_{0r}} = s, T_{i_{0r}} = 0) \\ & \cdot \log \mathbb{P}(Z_{i_{0r}} = s \mid Y_{i_{0r}} = j) \end{aligned}$$

and

$$\begin{aligned} l_{21}^{(t)}(\psi) = \sum_{j=0}^{k-1} \sum_{r=1}^{R_1} \sum_{s=j}^{k_Z-1+j} & n_1(r,s) \cdot \mathbb{P}_t(Y_{i_{1r}} = j \mid S_{i_{1r}} = s, T_{i_{1r}} = 1) \\ & \cdot \log \mathbb{P}(Z_{i_{1r}} = s - j \mid Y_{i_{1r}} = j). \end{aligned}$$

Notice, the probabilities $\mathbb{P}_t(Y_{i_{0r}} = j \mid S_{i_{0r}} = s, T_{i_{0r}} = 0)$ and $\mathbb{P}_t(Y_{i_{1r}} = j \mid S_{i_{1r}} = s, T_{i_{1r}} = 1)$ are already available from the calculation corresponding to $l_{10}^{(t)}(\beta)$ and $l_{11}^{(t)}(\beta)$.

In the M step of iteration $t+1$, we maximize $l_1^{(t)}$ and $l_2^{(t)} = l_{20}^{(t)} + l_{21}^{(t)}$ in β respectively ψ. The maxima are the new estimates $\beta_{(t+1)}$ and $\psi_{(t+1)}$. The vector $\beta_{(t+1)}$ is the MLE for an ordinary multivariate logistic regression model with the following data situation: There are $R_0 + R_1$ covariate levels. For covariate level equal to $x_{i_{0r}}$ $(r = 1, \ldots, R_0)$ the outcome $Y = j$ is observed

$$\left(\sum_{s=0}^{k_Z-1} n_0(r,s) \cdot \mathbb{P}_t(Y_{i_{0r}} = j \mid S_{i_{0r}} = s, T_{i_{0r}} = 0) \right)$$

times while for value $x_{i_{1r}}$ $(r = 1, \ldots, R_1)$ of the covariates the value $Y = j$ occurs

$$\left(\sum_{s=0}^{k_Z+k-2} n_1(r,s) \cdot \mathbb{P}_t(Y_{i_{1r}} = j \mid S_{i_{1r}} = s, T_{i_{1r}} = 1) \right)$$

times. Thus, one part of the data corresponds to the control group and the other part corresponds to the treatment group. Since we are working with a standard logistic regression situation (aside from the fact that noninteger observations appear), $\beta_{(t+1)}$ can be obtained with standard statistics software. The quantity $\psi_{(t+1)}$ can be computed similarly. Referring to this, note that $\psi_{(t+1)}$ is the MLE for a multivariate logistic regression model with data constellation as follows: The covariate levels for this constellation are given by $v(x_{i_{0r}}, j)$ as well as $v(x_{i_{1r}}, j)$ for $j = 0, \ldots, k-1$ and $r = 1, \ldots, R_0$ respectively $r = 1, \ldots, R_1$. I.e., the sensitive item can play the role of a covariate in this data setup. For the covariate level equal to $v(x_{i_{0r}}, j)$, the outcome $Z = s$ occurs $(n_0(r, s) \cdot \mathbb{P}_t(Y_{i_{0r}} = j \,|\, S_{i_{0r}} = s, T_{i_{0r}} = 0))$ times. For covariates equal to $v(x_{i_{1r}}, j)$, the value $Z = s$ appears

$$(n_1(r, j + s) \cdot \mathbb{P}_t(Y_{i_{1r}} = j \,|\, S_{i_{1r}} = j + s, T_{i_{1r}} = 1))$$

times ($s = 0, \ldots, k_Z - 1$). Due to this data constellation, $\psi_{(t+1)}$ can be calculated by standard software, too. After sufficiently many EM algorithm iterations, an estimate $(\hat{\beta}^\top, \hat{\psi}^\top)^\top$ is present.

Our next aim is a variance estimation for the estimator $(\hat{\beta}^\top, \hat{\psi}^\top)^\top$. For this goal, bootstrap resampling is again advantageous. We first remark that the probability of the event $\{S_i = s, T_i = t\}$ can be estimated by

$$\hat{\mathbb{P}}(S_i = s, T_i = t) = \sum_{j=0}^{k-1} \frac{1}{2} \cdot \hat{\mathbb{P}}(Z_i = s - t \cdot j \,|\, Y_i = j) \cdot \hat{\mathbb{P}}(Y_i = j) \qquad (19)$$

where $\hat{\mathbb{P}}(Z_i = s - t \cdot j \,|\, Y_i = j)$ is computed by replacing ψ by $\hat{\psi}$ in (15) and $\hat{\mathbb{P}}(Y_i = j) = \exp(x_i \hat{\beta}^{(j)})/(1 + \exp(x_i \hat{\beta}^{(1)}) + \cdots + \exp(x_i \hat{\beta}^{(k-1)}))$. We obtain the bth ($b = 1, \ldots, B$) bootstrap replication of $(\hat{\beta}^\top, \hat{\psi}^\top)^\top$ by drawing for $i = 1, \ldots, n$ a realization $(s_i^{(b)}, t_i^{(b)})$ according to (19) and employing the EM algorithm as described above to these new data. From the B resampled versions of $(\hat{\beta}^\top, \hat{\psi}^\top)^\top$, we can calculate an empirical variance matrix. This is the bootstrap estimate for the variance of $(\hat{\beta}^\top, \hat{\psi}^\top)^\top$. By calculating empirical quantiles from the replications, we obtain confidence intervals for the components of $(\beta^\top, \psi^\top)^\top$.

Let us now address stochastic covariates, that is, the values of the exogenous characteristics are random. The interview procedure is that the sample units report both the outcomes of the covariates and an answer according to the item count technique in Subsection 4.1. We introduce the random row vector X_i whose jth entry describes the ith respondents value of the jth covariate ($i = 1, \ldots, n$; $j = 1, \ldots, p$) and make the assumptions:

(S1') The n vectors $(Y_1, Z_1, T_1, X_1), \ldots, (Y_n, Z_n, T_n, X_n)$ are iid.

(S2') For every $i = 1, \ldots, n$, we have: T_i and (Y_i, Z_i, X_i) are independent and $\mathbb{P}(T_i = 1) = 1/2$.

(S3') A vector $\beta = (\beta^{(1)^\top}, \ldots, \beta^{(k-1)^\top})^\top$ with $\beta^{(j)} \in \mathbb{R}^{p\times 1}$ exists so that we have for $j = 1, \ldots, k - 1$

$$\mathbb{P}(Y_i = j \,|\, X_i = x) = \frac{e^{x\beta^{(j)}}}{1 + e^{x\beta^{(1)}} + \cdots + e^{x\beta^{(k-1)}}}.$$

Consequently, we are in the situation of a multivariate logistic regression model with stochastic covariates. (S1') and (S2') hold when we apply simple random sampling with replacement for generating the sample and assign each sample unit to the control or treatment group by e.g. tossing a fair coin. For stochastic covariates the model (15) for the control items now changes to

$$\mathbb{P}(Z_i = z \,|\, Y_i = y, X_i = x) = \frac{e^{v(x,y)\cdot\psi^{(z)}}}{1 + e^{v(x,y)\cdot\psi^{(1)}} + \cdots + e^{v(x,y)\cdot\psi^{(k_Z-1)}}}.$$

Then, the observed data log-likelihood is $l_{obs}(\beta, \psi) =$

$$\sum_{i=1}^{n} \log \left[\sum_{j=0}^{k-1} \mathbb{P}(Z_i = s_i - t_i j \,|\, Y_i = j, X_i = x_i) \cdot \mathbb{P}(Y_i = j \,|\, X_i = x_i) \right] \tag{20}$$

where s_i, t_i, and x_i are the observed realizations of S_i, T_i, and X_i, respectively. This log-likelihood has the same form as (16). Thus, maximizing (20) is equivalent to the maximization of a log-likelihood that corresponds to the deterministic case. In other words, we can trace the ML estimation for stochastic exogenous variables back to the ML estimation for deterministic covariates. We can obtain the estimator's variance given the observed covariates by a bootstrap resampling method that proceeds analog to the case of deterministic covariates.

4.5. An ICT for polychotomous variables without control group

In Section 3, we have studied an ICT for binary variables without control group while for the polychotomous case, the presence of a control group was required so far in Subsections 4.1 - 4.4. In this subsection, we show that it is possible to dispense with the control group also in the polychotomous case. In the setup without control group, every respondent is instructed to give answer $Y + Z$ and we have to make assumptions analog to (2) and (3), that is, we assume that the distribution of Z is known and that Z and Y are independent. The observed data log-likelihood is then given by

$$l_{obs}(\pi; \mathbf{s}) = \sum_{i=1}^{n} \log \mathbb{P}(S_i = s_i) = \sum_{i=1}^{n} \log \left[\sum_{j=0}^{k-1} \phi_{s_i - j} \cdot \pi_j \right]. \tag{21}$$

To maximize (21), we apply the EM algorithm again where we must adapt the calculation from Subsection 4.1 as follows. The complete data log-likelihood is

$$l_{com}(\pi) = l_{com}(\pi; \mathbf{y}, \mathbf{z}, \mathbf{s}) = \sum_{i=1}^{n} \log \mathbb{P}(Y_i = y_i) = \sum_{j=0}^{k-1} \log \pi_j \cdot \sum_{i=1}^{n} 1_{\{j\}}(y_i).$$

In the E step of the EM algorithm in iteration $t + 1$, we compute

$$\begin{aligned}\widehat{l_{com}}(\pi) &= \mathbb{E}_t(l_{com}(\pi; \mathbf{Y}, \mathbf{Z}, \mathbf{S}) |\, \mathbf{S} = \mathbf{s}) \\ &= \sum_{j=0}^{k-1} \log \pi_j \cdot \sum_{i=1}^{n} \mathbb{E}_t(1_{\{j\}}(Y_i) | \mathbf{S} = \mathbf{s}) =: \sum_{j=0}^{k-1} \log \pi_j \cdot v_j^{(t)},\end{aligned}$$

where

$$\mathbb{E}_t(1_{\{j\}}(Y_i) | \mathbf{S} = \mathbf{s}) = \mathbb{P}_t(Y_i = j | S_i = s_i) = \frac{\phi_{s_i - j} \cdot \pi_j^{(t)}}{\sum_{l=0}^{k-1} \phi_{s_i - l} \cdot \pi_l^{(t)}}$$

holds. In the M step of iteration $t+1$, we obtain $\pi_j^{(t+1)} = v_j^{(t)}/(v_0^{(t)} + \cdots + v_{k-1}^{(t)})$. It is easy to see that the maximization of (21) is equivalent to the maximization of $\tilde{l}$, which possesses $k-1$ arguments and is given by

$$\begin{aligned}\tilde{l}(\pi_0, \ldots, \pi_{k-2}) &= n_1 \cdot \log[\boldsymbol{\phi} \cdot (\pi_0, \ldots, \pi_{k-2}, 1 - \pi_0 - \cdots - \pi_{k-2})^\top] \\ &= n_1 \cdot \log[\boldsymbol{\phi}^* \cdot (\pi_0, \ldots, \pi_{k-2})^\top + \phi^*].\end{aligned}$$

Here, log is applied componentwise, $\boldsymbol{\phi}$ is a $(k_Z + k - 1) \times k$ matrix whose entry (i, j) is ϕ_{i-j} for $i = 0, \ldots, k_Z + k - 2$ and $j = 0, \ldots, k-1$, n_{1i} represents the absolute frequency of answer i in the sample, and $n_1 = (n_{10}, \ldots, n_{1,k_Z+k-2})$. Moreover, the jth column of $\boldsymbol{\phi}^*$ $(j = 0, \ldots, k-2)$ equals the jth column of $\boldsymbol{\phi}$ minus the last column of $\boldsymbol{\phi}$ and ϕ^* is equal to the last column of $\boldsymbol{\phi}$. Some standard calculations deliver the Hessian matrix

$$\tilde{l}''(\pi_0, \ldots, \pi_{k-2}) = [\boldsymbol{\phi}^*]^\top \cdot (-D) \cdot \boldsymbol{\phi}^* = \quad - \quad [\boldsymbol{\phi}^*]^\top \cdot \sqrt{D} \cdot \sqrt{D} \cdot \boldsymbol{\phi}^*$$

with a diagonal matrix D whose ith diagonal element $(i = 0, \ldots, k_Z + k - 2)$ equals n_{1i}/λ_i^2 where $(\lambda_0, \ldots, \lambda_{k_Z+k-2})^\top = \boldsymbol{\phi} \cdot (\pi_0, \ldots, \pi_{k-2}, 1 - \pi_0 - \cdots - \pi_{k-2})^\top$. Assuming strict positivity of the ϕ_i and n_{1i} (i.e., $\phi_i > 0$ and $n_{1i} > 0$), one can show that the Hessian matrix is negative definite implying that $\tilde{l}$ is strictly concave. Furthermore, the Fisher matrix is given by

$$\begin{aligned}F = F(\pi_0, \ldots, \pi_{k-2}) &= \mathbb{E}(-\tilde{l}''(\pi_0, \ldots, \pi_{k-2}; S_1, \ldots, S_n) \\ &= n \cdot [\boldsymbol{\phi}^*]^\top \cdot diag(\frac{1}{\lambda_0}, \ldots, \frac{1}{\lambda_{k_Z+k-2}}) \cdot \boldsymbol{\phi}^*\end{aligned}$$

and the asymptotic variance of the ML estimator $(\hat{\pi}_0, \ldots, \hat{\pi}_{k-2})$ is given by the inverse of F. Of course, we also can compute bootstrap variance estimates, however, we skip details here, because the procedure should be clear from previous parts of this paper. The above descriptions show how the computations for the ML estimation of unconditional proportions from previous parts of this paper have to be adapted when a multicategorical sensitive characteristic is intended to be investigated with an ICT without control group. Given these detailed descriptions, the adaption of the explanations on Bayes and regression estimation from previous parts of this article on the case of a multichotomous variable and an ICT without control group is quite straightforward. For this reason, we do not go into particulars here.

5. Concluding remarks

When data on sensitive topics are intended to be collected in a survey, direct questions such as “Have you ever committed tax evasion?” are not advisable. The reason is that they lead to missing values due to answer refusal or untruthful responses. Hence, ingenious procedures for the survey are necessary. One such approach is the item count technique. The basic principle of the ICT is that only the overall sum of outcomes of a sensitive characteristic and several innocuous characteristics is revealed. This idea usually protects the privacy. Thus, we can expect the ICT to deliver more trustworthy estimates than direct questioning.

To gather data on sensitive attributes, different alternatives to the ICT are available in the literature. One of these is the nonrandomized response (NRR) approach (see e.g. Tian and Tang (2014, [19])). In NRR schemes, the desired scrambled answer is a function of the sensitive variable and a nonsensitive scrambling variable. Moreover, every respondent gives the same answer if he or she is interviewed repeatedly. In fact, these features of NRR methods hold also for the item count techniques without control group from Section 3 and Subsection 4.5. In particular, Z plays the role of a scrambling variable while the scrambled answer is $S = Y + Z$. Thus, such an ICT can be considered as a special NRR technique. A further approach for gathering sensitive data are randomized response (RR) techniques (e.g., Chaudhuri (2011, [3])). In comparison with the ICT, these methods possess, however, the uncomfortable feature that the respondents must accomplish a random experiment with the help of a randomization device. We further remark that our item count techniques from Section 3 and Subsection 4.5 show similarities to known additive data masking techniques. Additive data masking was first investigated by Pollock and Bek (1976, [16]), whose work was extended in several directions (e.g., in Gupta et al. (2010, [10]), some respondents have the option to provide a direct answer; Eichhorn and Hayre (1983, [8]) and Bar-Lev et al. (2004, [1]) consider multiplicative masking). The basic idea in Pollock and Bek (1976, [16]), who have quantitative sensitive variables in mind, is that the interviewee adds the value of a random number from a known distribution to the value of the sensitive characteristic. This answer principle resembles the answer principle of the methods in Section 3 and Subsection 4.5. However, the estimations given in Pollock and Bek (1976, [16]) are not useful for our constellations, because their moment estimator can deliver inadmissible values and their parametric estimation is only briefly mentioned and only illustrated in a very simple example. Moreover, our techniques can be classified as NRR method whereas the technique from Pollock and Bek (1976, [16]) is a RR procedure.

Several studies demonstrate that the ICT approach can be successful to gather sensitive data (e.g., Tsuchiya et al. (2007, [22]), Holbrook and Krosnick (2010, [11]), Coutts and Jann (2011, [6]), and Trappmann et al. (2014, [20])). Moreover, a number of useful estimation methods regarding the ICT have been developed in recent years. Nevertheless, several methodological gaps remained so far. Important gaps are addressed in this paper. In particular, we have described a generalized ICT for binary attributes without control group and derived admissible estimators, presented Bayesian inference and established methods for regression analysis. Furthermore, we have considered the field of multicategorical sensitive characteristics. Here, we have derived a version of the ICT for such attributes including unconditional MLEs, Bayes estimates, and regression estimates.

The item count methods presented in this paper inherit the problem that certain answers restrict the possible values of the sensitive variable (ceiling / floor effect) from the originally ICT by Miller (1984, [14]). As described, suitable chosen control items can partially mitigate this problem. One possible starting point for future research is to perfect the privacy protection offered by the ICT. For instance, the ICT from Section 3 could be modified such that respondents who originally have to give the maximal answer $S = J + 1$ should give another predefined response, e.g., response 0. Then, the ceiling effect is avoided. Such a revised answer scheme was already briefly mentioned in Petroczi (2011, p. 11, [15]), but without presenting the corresponding estimation. The complete development of ML, Bayes, and regression estimates for this modified ICT is beyond the scope of our paper, too. Nevertheless, we outline the ML estimation: First, note that the modified procedure results in new probabilities for the answers (the possible answers are now $0, \ldots, J$). These probabilities can be easily calculated by thinking about which combinations of Y and Z lead to a certain answer. With these new probabilities, we obtain an observed data log-likelihood function, which can be maximized similar to Section 3 via the EM algorithm.

Acknowledgements

The author would like to thank Professor Dr. Karlheinz Fleischer, two referees, and the Editor for their comments and suggestions, which helped to improve the article.

References

[1] Bar-Lev, S.K., Bobovitch, E. and Boukai, B. (2004). A note on randomized response models for quantitative data. Metrika 60, 255–260. MR2189754

[2] Blair, G. and Imai, K. (2012). Statistical analysis of list experiments. Political Analysis 20, 47–77.

[3] Chaudhuri, A. (2011). Randomized Response and Indirect Questioning Techniques in Surveys. Chapman & Hall/CRC. MR2759226

[4] Chaudhuri, A. and Christofides, T.C. (2007). Item count technique in estimating the proportion of people with a sensitive feature. Journal of Statistical Planning and Inference 137, 589–593. MR2298959

[5] Chaudhuri, A. and Christofides, T.C. (2013). Indirect Questioning in Sample Surveys. Springer. MR3100051

[6] Coutts, E. and Jann, B. (2011). Sensitive questions in online surveys: Experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). Sociological Methods & Research 40, 169–193. MR2758303

[7] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B 39, 1–38. MR0501537

[8] Eichhorn, B.H. and Hayre, L.S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. Journal of Statistical Planning and Inference 7, 307–316.

[9] Groenitz, H. (2013). Using prior information in privacy-protecting survey designs for categorical sensitive variables. Statistical Papers, DOI 10.1007/s00362-013-0573-3.

[10] Gupta, S., Shabbir, J. and Sehra, S. (2010). Mean and sensitivity estimation in optional randomized response models. Journal of Statistical Planning and Inference 140, 2870–2874. MR2651974

[11] Holbrook, A.L. and Krosnick, J.A. (2010). Social desirability bias in voter turnout reports: Tests using the item count technique. Public Opinion Quarterly 74, 37–67.

[12] Imai, K. (2011). Multivariate regression analysis for the item count technique. Journal of the American Statistical Association 106, 407–416. MR2866971

[13] Kuha, J. and Jackson, J. (2014). The item count method for sensitive survey questions: Modelling criminal behaviour. Journal of the Royal Statistical Society C 63, 321–341. MR3234346

[14] Miller, J.D. (1984). A New Survey Technique for Studying Deviant Behavior. PhD thesis, The George Washington University.

[15] Petroczi, A., Nepusz, T., Cross, P., Taft, H., Shah, S., Deshmukh, N., Schaffer, J., Shane, M., Adesanwo, C., Barker, J. and 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.

[16] Pollock, K.H. and Bek, Y. (1976). A comparison of three randomized response models for quantitative data. Journal of the American Statistical Association 71, 884–886.

[17] Schafer, J.L. (2000). Analysis of Incomplete Multivariate Data. Chapman & Hall/CRC. MR1692799

[18] Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82, 528–540. MR0898357

[19] Tian, G.L. and Tang, M.L. (2014). Incomplete categorical data design: Non-randomized response techniques for sensitive questions in surveys. CRC Press. MR3114330

[20] Trappmann, M., Krumpal, I., Kirchner, A. and Jann, B. (2014). Item sum: A new technique for asking quantitative sensitive questions. Journal of Survey Statistics and Methodology 2, 58–77.

[21] Tsuchiya, T. (2005). Domain estimators for the item count technique. Survey Methodology 31, 41–51.

[22] Tsuchiya, T., Hirai, Y. and Ono, S. (2007). A study of the properties of the item count technique. Public Opinion Quarterly 71, 253–272.