Research Article

Statistics in Medicine

Received 1 March 2012, Accepted 29 April 2013 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5859

Mixture randomized item-response modeling: a smoking behavior validation study

J.-P. Fox,[a*†] M. Avetisyan[a] and J. van der Palen[a,b]

Misleading response behavior is expected in medical settings where incriminating behavior is negatively related to the recovery from a disease. In the present study, lung patients feel social and professional pressure concerning smoking and experience questions about smoking behavior as sensitive and tend to conceal embarrassing or threatening information. The randomized item-response survey method is expected to improve the accuracy of self-reports as individual item responses are masked and only randomized item responses are observed.

We explored the validation of the randomized item-response technique in a unique experimental study. Therefore, we administered a new multi-item measure assessing smoking behavior by using a treatment–control design (randomized response (RR) or direct questioning). After the questionnaire, we administered a breath test by using a carbon monoxide (CO) monitor to determine the smoking status of the patient. We used the response data to measure the individual smoking behavior by using a mixture item-response model. It is shown that the detected smokers scored significantly higher in the RR condition compared with the directly questioned condition. We proposed a Bayesian latent variable framework to evaluate the diagnostic test accuracy of the questionnaire using the randomized-response technique, which is based on the posterior densities of the subject's smoking behavior scores together with the breath test measurements. For different diagnostic test thresholds, we obtained moderate posterior mean estimates of sensitivity and specificity by observing a limited number of discrete randomized item responses.

Keywords: randomized response; validation; mixture item response theory; diagnostic test accuracy; classification probabilities

1. Introduction

The research on smoking status assessment often rests on information available from self-reports. Many researchers indicate that denial and underreporting of the extent of smoking are quite usual in certain populations, such as adolescents and announced quitters [1, 2]. Self-reports are also known to be less reliable for special subgroups such as individuals with a coronary diagnosis or other smoking-related diseases [3]. This follows from the fact that these smokers fail to discontinue smoking but feel a strong pressure to do so. However, for a wide variety of medical conditions and particularly for diseases of the respiratory system, accurate information about the smoking status of patients is crucial for the choice of suitable treatments. In addition, accurate information is also essential for healthcare professionals during smoking cessation programs. In general, survey research techniques have several limitations that can undermine its usefulness in health research. One of the limitations is that individuals are reluctant to provide personal information about sensitive attitudes or behaviors [4]. Respondents tend to under-report or over-report or avoid questions that are perceived as threatening or sensitive. Threatening or sensitive questions can concern personal behavior (such as sexual practices), illegal behavior (such as drug use or alcohol consumption), and other health-related behaviors (such as smoking).

[a] *Department of Research Methodology, Measurement, and Data Analysis, Faculty of Behavioral Sciences, University of Twente, Enschede, The Netherlands*

[b] *Department of Pulmonary Medicine, Medisch Spectrum Twente, Enschede, The Netherlands*

**Correspondence to: J.-P. Fox, Department of Research Methodology, Measurement, and Data Analysis, Faculty of Behavioral Sciences, University of Twente, The Netherlands.*

[†] *E-mail: j.p.fox@utwente.nl*

Many agree on the fact that a validation of self-reported levels of smoking is necessary [5, 6]. A broad spectrum of biochemical validation techniques that differentiate between smokers and non-smokers is available in clinical settings. Many methods are based on the determination of the levels of nicotine and its metabolites in blood or urine, as well as the traces of carbon monoxide (CO) in blood and saliva [7]. Most of these tests are rather invasive, expensive, or time consuming. An alternative, less expensive method offering immediate result, not requiring specialized training, and suitable for situations where nicotine replacement strategies are used is the expired-air CO test.

Attention will be focused on the improvement of self-reports on a sensitive attribute by means of alternative survey techniques, ensuring the privacy of the respondents' answers. In particular, Warner [8] proposed a univariate randomized response (RR) data collection method that led to a whole family of techniques, which insure confidentiality of responses due to randomization. For example, in the so-called forced RR design, a randomizing device is used before an answer is given, and the outcome decides whether a truthful or forced (simulated) response is requested. The answer is protected as the outcome of the randomizing device is only known to the respondent.

Although Soeken [9], Rittenhouse [10, 11], and Williams *et al.* [12] explicitly recommended the use of the RR technique (RRT) in health research, the RR survey method did not receive much attention. Only recently, Cross *et al.* [13] used the RRT to estimate the sheep scab prevalence in Welsh flocks by interviewing farmers in Wales, and Ostapczuk *et al.* [14] used the RRT to investigate lifetime prevalence of medication non-adherence in Germany.

The little attention toward the univariate RRTs in health research may be attributed to the limitation of inferences to the population level. However, it is possible to make individual-level inferences given multivariate RR data [15]. In other research fields, the RRT has obtained more attention, and various studies have been performed successfully to measure the prevalence of a sensitive behavior. De Jong *et al.* [16, 17] reported on desires for products and services in the domain of adult entertainment. Fox and Meijer [18] and Fox [15] analyzed students' academic cheating behavior at a Dutch university. Fox and Wyrick [19] measured the prevalence of alcohol use and alcohol-related problems among college students. Rates of academic cheating and rates of criminal behavior were investigated by Tracy and Fox [20] and Scheers and Dayton [21], respectively. Lensvelt-Mulders *et al.* [22, 23] and van der Heijden *et al.* [24] discussed the prevalence of academic cheating, tax evasion, and software piracy, among others.

In the present experimental smoking behavior survey study, we validated the RRT by comparing the survey results with the CO-test outcomes. Therefore, we randomly selected participants into two groups, where we questioned one group (RR group) using the forced randomized-response technique and we directly questioned (DQ) the other group (DQ group). We collected response data using a new multi-item smoking scale questionnaire, comprising ordinal and dichotomous items to measure smoking behavior. We determined the true smoking status (smoker/non-smoker) of every participant after the interview by the expired-air CO test, which served as the gold standard.

An item response theory model [25] combined with an RR model will be used to measure the sensitive latent trait smoking behavior given multivariate RR data. The so-called randomized item-response model enables the measurement of item and person characteristics, while accounting for the RR nature of the data being analyzed. Within a Bayesian framework, we estimated individual latent scores on the smoking behavior scale. At the level of the individual, we defined a mean structure for the smoking behavior variable that includes the random assignment of subjects to treatment and control groups and various explanatory background variables. To validate the RRT, we explained latent score differences through a regression analysis by individual, group, questioning technique differences, and the true smoking status from the CO measurement.

In practice, the RR design may not always be followed. Non-compliant respondents (cheaters) are expected to elicit the least incriminating response to improve their self-report. This leads to a surplus of such responses. Therefore, Clark and Desharnais [26], Böckenholt and van der Heijden [27], De Jong *et al.* [16], and Ostapczuk *et al.* [14], defined a mixture RR model to account for respondents that do not follow the RR instructions. A generalized mixture model is proposed using response-type specific (binary and ordinal) latent classes to capture extreme use of the less incriminating response category. The developed mixture randomized item-response model for mixed response types (ordinal and binary) generalizes the models of Böckenholt and van der Heijden [24], De Jong *et al.* [16], and Fox [15, 28]. The generalization makes it possible to model heterogeneity in response-type specific positive self-presentation biases. It is expected that the binary response format will reveal bias (positive self-presentation) more clearly than the ordinal response format. Multiple ordered response categories offer

respondents more degrees to express their behavior compared with the binary response range and allow respondents to misreport their true behavior in varying degrees of self-protective response behavior.

To further evaluate the performance of the RRT, we compared the diagnoses (smoker/non-smoker), following from the test results, with the true smoking status (i.e., CO test results). The diagnoses cannot be observed directly. However, the continuous latent level of smoking behavior can be translated to a categorical latent smoking status (referred to as the diagnosis) given a selected decision threshold on the continuous latent scale. To assess the diagnostic test accuracy of the smoking questionnaire under RR questioning, we proposed expected posterior classification probabilities, such as the expected posterior sensitivity and specificity, where the expectation is taken over the posterior distribution of the diagnoses. This makes it possible to evaluate the diagnostic accuracy of the smoking questionnaire given the true smoking status and the posterior distribution of test diagnoses, such that the uncertainty in the test diagnoses is taken into account. In the same way, we further validated the RRT using the predictive properties of the smoking behavior questionnaire. That is, the positive and negative predictive values of the test under the randomized item-response technique will be analyzed.

This paper is organized as follows. First, we presented the forced RR method and the study design. Thereafter, we discussed the mixture randomized item-response model for mixed responses. We described the Bayesian latent variable method for diagnostic accuracy together with expected posterior classification probabilities. Then we used the model and the Bayesian test diagnostics to validate the RRT and to examine the properties of the smoking behavior test in an experimental-clinical smoking study. Finally, we give some concluding comments and suggestions for further research.

2. Method

This study involved outpatients of a pulmonary department of a hospital in the Netherlands. Healthcare providers strongly recommend lung patients to quit smoking; therefore, the patients that are not complying with the medical advice often experience questions about smoking as very sensitive.

Shadel and Shiffman [29] gave an extensive overview of methods of assessment of smoking behavior and assessed aspects such as amount of smoking, nicotine dependence, and withdrawal symptoms. Although various scales in the smoking assessment literature are discussed, recommendations and guidance on which scale should be used to assess a particular construct are missing. A well-known short self-report measure is the Fagerström Tolerance Questionnaire (FTQ; [30]), but it is purely designed to assess the level of nicotine dependence. Other more specific tests are focused on smoking patterns and smoking urges and are most often used in smoking cessation programs. Therefore, to measure the general construct smoking behavior, a multi-item questionnaire was developed using items that cover different aspects of smoking behavior as reported in Shadel and Shiffman [29].

The items comprise history of smoking, compulsion to smoke and craving, influences of other factors, and consequences of smoking. As most clinical screening instruments about smoking, items are used to assess the smoking history including whether the subject ever smoked, for how long, and the smoking frequency (e.g., smoking patterns). Items are selected that focus on the compulsion to smoke and craving, which relates to the nicotine dependence syndrome. Furthermore, items are selected that cover smoking behavior influenced by other factors and consequences of smoking. The items also cover questions that are usually asked by medical personal during the routine visits to the treating pulmonologist. This multi-item questionnaire (Appendix B) comprises nine dichotomous and three polytomous items and was used to assess subject's smoking behavior.

According to a randomized control trial, we randomly assigned subjects to an RR or DQ condition. In the DQ condition, subjects completed the questionnaire in a conventional way. In the RR condition, subjects completed the questionnaire using a randomizing spinner device for the items with two and three response categories.

We recorded the demographic characteristics such as gender, age, and educational level of the patient as background characteristics, which are known to be related to smoking behavior [31]. In addition, we also noted information on medical condition and treating pulmonologist.

After completion of the questionnaire, we assessed the smoking status of the participants by means of physical CO level measurement in the expired air [32]. We used the Bedfont Micro 4 (Bedfont Scientific Ltd Station Road Harrietsham Maidstone Kent ME17 1JA England) Smokerlyzer portable CO monitor for measuring expired-air CO level. We used the CO measurement to distinguish smokers from non-smokers, which was referred to as the true smoking status of the patient. We used the response data to estimate patient's smoking behavior, as a continuous latent trait, and patient's smoking status,

as a discrete latent trait. Subsequently, we used the true smoking status to validate RR as a questioning technique to improve the accuracy of self-reports.

2.1. The randomized-response technique

In the RR condition, to answer an item, a respondent spun a spinner and was instructed to give a forced response or an honest response. That is, the outcome of the spinner determined whether the respondent was requested to answer honestly or to give a (simulated) forced response. Respondent's answers were protected, as outcomes of the spinner were only known to the respondents. We explained this protection mechanism at the start of the interview such that respondents were stimulated to provide honest answers. Each observed individual response could be a forced response, which made it much easier for the individuals to give truthful responses to sensitive questions when an honest response was requested by the randomizing device.

To illustrate the procedure, consider the randomizing device for the binary items. This randomizing device is a spinner with 18 equal sectors, where 14 sectors are marked H (honest response), two sectors marked Y (forced 'yes' response), and two sectors marked N (forced 'no' response) as possible outcomes. When a respondent spins the spinner, the probability of the outcome H is 14 out of 18, and with probability $\phi_1 = 14/18$, an honest response is requested. Four sectors represent a forced response, and with probability $(1-\phi_1) = 4/18$, the respondent is instructed to give a forced response. A forced yes or no response is each instructed in two out of four possible outcomes such that $\phi_2 = 0.50$. Thus, when the arrow is spun, it can land in a sector marked H, and an honest response is requested, but it can also land in sector Y or N, and a forced yes or no response is requested, respectively.

We constructed the spinner for items with three response options such that an honest response was requested with probability 0.611. When a forced response was dictated, the instruction was to give a response in category one, two, or three with probability 0.25, 0.50, and 0.25, respectively.

2.2. Mixture randomized item-response model

According to the forced RR design discussed, an honest response is requested with probability ϕ_1 and a forced response with probability $(1-\phi_1)$. Subsequently, a forced positive response is prompted with probability ϕ_2. The a priori known characteristics of the randomizing device determine the probabilities ϕ_1 and ϕ_2. Let Y_{ik} denote the RR of subject i to item k. Subsequently, let $\tilde{Y}_{ik}$ denote the response that is observed when an honest response is requested, which is referred to as the non-randomized response. The probability of an observed RR given by participant i to item k can be expressed as

$$P\left(Y_{ik}=c\right)=\phi_1 P\left(\tilde{Y}_{ik}=c\right)+(1-\phi_1)\phi_2(c)\; c\in 1,\ldots,C. \tag{1}$$

We modeled the observed RR data using a mixture randomized item-response model for mixed responses (binary and ordinal). For each response-type format, we assumed that the mixture component is represented by two unobserved groups in the population. The subjects belong to the compliance group who follow the randomization scheme or to the non-compliance group who always choose the least stigmatizing category. This non-compliance group is characterized by subjects that respond negatively (the least self-incriminating answer) with probability one to all questions and ignore the instructions of the RR design. We assumed the randomized item responses of subjects in the compliance group to be distributed according to an item-response model when an honest answer is requested and a generated (forced) response when a forced response is prompted by the randomizing device. We assumed the responses from the DQ group to be distributed according to an item-response model. Furthermore, we assumed the item characteristics to be invariant over questioning techniques.

Let θ_i denote the latent smoking behavior of subject i, and b_k the threshold parameter of item k. For subject i $(i=1,\ldots,N)$, the probability of a positive response to item k depends on the random assignment to the RR or DQ group and its membership to the compliance or non-compliance group. When item k has two response categories, it is shown in Appendix A that the response probabilities of subject i to item k are given by

$$\begin{aligned}
P\left(Y_{ik}=0\mid\theta_i,b_k\right)&=\begin{cases}1-\Phi\left(\theta_i-b_k\right) & \text{for DQ}\\ \pi_i^b+\left(1-\pi_i^b\right)\left(1-\left(\phi_1\Phi\left(\theta_i-b_k\right)+(1-\phi_1)\phi_2\right)\right) & \text{for RR,}\end{cases}\\
P\left(Y_{ik}=1\mid\theta_i,b_k\right)&=\begin{cases}\Phi\left(\theta_i-b_k\right) & \text{for DQ}\\ \left(1-\pi_i^b\right)\left(\phi_1\Phi\left(\theta_i-b_k\right)+(1-\phi_1)\phi_2\right) & \text{for RR,}\end{cases}
\end{aligned} \tag{2}$$

where $\Phi(.)$ denotes the cumulative normal distribution function. For the binary items, $\pi_i^b = 1$ when subject i belongs to the non-compliance group and $\pi_i^b = 0$ otherwise.

When item k has three response categories, the response probabilities of subject i are given by

$$\begin{aligned}
P\left(Y_{ik}=0 \mid \theta_i, \kappa_{k1}\right) &= \begin{cases} \Phi\left(\kappa_{k1}-\theta_i\right) & \text{for DQ} \\ \pi_i^o + \left(1-\pi_i^o\right)\left(\phi_1 \Phi\left(\kappa_{k1}-\theta_i\right) + (1-\phi_1)\phi_2(0)\right) & \text{for RR} \end{cases} \\
P\left(Y_{ik}=1 \mid \theta_i, \kappa_k\right) &= \begin{cases} \Phi\left(\theta_i-\kappa_{k1}\right) - \Phi\left(\theta_i-\kappa_{k2}\right) & \text{for DQ} \\ \left(1-\pi_i^o\right)\left(\phi_1\left[\Phi\left(\theta_i-\kappa_{k1}\right) - \Phi\left(\theta_i-\kappa_{k2}\right)\right] + (1-\phi_1)\phi_2(1)\right) & \text{for RR} \end{cases} \\
P\left(Y_{ik}=2 \mid \theta_i, \kappa_{k2}\right) &= \begin{cases} \Phi\left(\theta_i-\kappa_{k2}\right) & \text{for DQ} \\ \left(1-\pi_i^o\right)\left(\phi_1 \Phi\left(\theta_i-\kappa_{k2}\right) + (1-\phi_1)\phi_2(2)\right) & \text{for RR,} \end{cases}
\end{aligned} \tag{3}$$

where, for the ordinal items, $\pi_i^o = 1$ when subject i belongs to the non-compliance group and $\pi_i^o = 0$ otherwise.

At a higher level, general priors for the model parameters are specified as follows:

$$\begin{aligned}
\theta_i &\sim \mathcal{N}\left(\mu_\theta, \sigma_\theta^2\right) \\
b_k &\sim \mathcal{N}\left(\mu_b, \sigma_b^2\right) \\
\pi_i^b &\sim \mathcal{B}\left(\pi_{01}\right) \\
\pi_i^o &\sim \mathcal{B}\left(\pi_{02}\right).
\end{aligned}$$

Furthermore, restricted normally distributed priors are specified for the thresholds κ_{k1} and κ_{k2} such that $\kappa_{k1} < \kappa_{k2}$. The hyperpriors are specified as follows:

$$\begin{aligned}
p\left(\mu_\theta \mid \sigma_\theta^2\right) &= \mathcal{N}\left(0, \sigma_\theta^2/n_0\right) \\
p\left(\mu_b \mid \sigma_b^2\right) &= \mathcal{N}\left(0, \sigma_b^2/n_0\right) \\
p\left(\pi_{01}\right) = p\left(\pi_{02}\right) &= \mathcal{B}e\left(g_1, g_2\right)
\end{aligned}$$

and inverse-gamma priors for the variances σ_b^2 and σ_θ^2.

2.2.1. Linear regression structure on smoking behavior. The normal population prior for the latent variable smoking behavior is modified to model mean latent differences between questioning techniques, to include effects of background information, and/or to analyze the relationship between the self-report measure of smoking behavior and the CO measure of smoking. Therefore, consider explanatory indicator variables *RR* and *CO*. The *RR* variable takes the value 1 when subjects are assigned to the *RR* condition (i.e., $RR = 1$) and 0 otherwise. Variable *CO* takes the value 1 when subjects are detected as smokers according to the *CO* test (i.e., $CO = 1$) and 0 otherwise. Now, the following linear regression on smoking behavior variable will be of specific interest, where both indicator variables and their interaction are included;

$$\theta_i = \beta_0 + \beta_1 RR_i + \beta_2 CO_i + \beta_3 RR_i CO_i + \epsilon_i, \tag{4}$$

where the intercept, β_0, represents the mean of non-smokers ($CO = 0$) in the DQ group, and β_1 and β_2 represent the effect of the randomized-response technique and of CO-measured smokers, respectively. Independent normally distributed error terms will be assumed such that $\epsilon_i \sim \mathcal{N}\left(0, \sigma_\theta^2\right)$.

The interaction effect, β_3, denotes the effect of CO-measured smokers in the RR group. The interaction effect will be of specific interest, as it can be used to validate the RRT. In the present study, this interaction effect will reveal whether identified smokers (according to the CO measurement) score significantly higher, when their responses are protected according to the randomized-response technique. Tate and Schmitz [33] validated a revised FTQ by investigating correlations between CO values and total-test scores, which were measured at the same time. However, they did not correct correlations for relevant background differences, and they did not take into account heterogeneity in item thresholds in the computation of the total scores. In the present nonlinear regression approach, we investigated the relation between CO measurements and smoking behavior scores while accounting for experimental differences, item threshold differences, and the discrete nature of item scores.

The mixture randomized item-response model parameters can be estimated simultaneously using MCMC. The WinBUGS program code is given in the Appendix C. We identified the model by restricting directly the mean of the latent variable or through a tight prior on the intercept. In both cases, we identified the mean of the latent scale. We simultaneously analyzed the direct questioning and RR data. As a result, we defined a common scale for the subjects in the RR and DQ conditions.

2.3. *Bayesian latent variable methods for diagnostic accuracy*

The outcomes of the expired-air CO test, hereafter referred to as the reference test, reveals the true smoking status of every participant. To validate the randomized-response technique, we compared measurements of smoking behavior based on randomized-response data and on direct-questioning data with the true smoking status. To investigate whether valid inferences can be made from the self-reports under RR and direct questioning, we compared diagnoses (smoker/non-smoker) on the basis of the self-reports and the CO measurements by using the well-known diagnostic accuracy measures. This takes the procedure of Heatherton *et al.* [30] one step further. They validated the Fagerström test for nicotine dependence by predicting biochemical measures (CO and cotinine measurements) using total-test scores through regression analyzes.

To make a comparison with the true smoking status according to the reference test, we assessed the smoking status of each subject. The smoking behavior measurement is considered to be a more general measurement, which includes the smoking status. When the posterior probability that the smoking behavior score is above a specific threshold is significantly high, the smoking status is positively diagnosed, and a positive diagnosis of smoking status of subject i will be denoted as $X_i = 1$, where the true smoking status will be denoted as $D_i = 1$. Consider a threshold value on the latent scale, θ_c; the posterior probability of a positive diagnosis $X_i = 1$, conditional on the response pattern, is given by

$$\begin{aligned} p_{ic} &= P\left(X_i = 1 \mid \mathbf{y}_i, \theta_c\right) = P\left(\theta_i > \theta_c \mid \mathbf{y}, \theta_c\right) \\ &= \int_{-\infty}^{\theta_c} p\left(\theta_i \mid \mathbf{y}\right) \mathrm{d}\theta_i . \end{aligned} \tag{5}$$

The subject's smoking status is a latent variable, and its measurement is derived from the subject's posterior density of smoking behavior. It follows that the null hypothesis stated as the smoking status of subject i equals 0 (non-smoker) is rejected when the posterior probability p_{ic} is higher than the significance level.

Garrett *et al.* [34] and Formann and Kohlmann [35] evaluated subject-specific medical diagnosis via a latent class analysis. In their approach, multiple test indicators are observed, and the posterior probability of having the disease given the test results is computed via a latent class analysis. In such a two-component latent class approach, the estimated subject-specific latent class probabilities may depend on the other members in the group. In the present approach, we used the posterior distribution of the latent variable smoking behavior to assess the smoking status (Equation (5)). In a two-step way, we used the subject-specific response data to measure smoking behavior, and we used the posterior distribution of smoking behavior to assess the smoking status (smokers and non-smokers).

2.3.1. Diagnostic accuracy given observed test diagnoses. In a more formal way, we compared the assessed smoking status or diagnoses with the true smoking status, and we evaluated the diagnostic accuracy by computing the operating characteristics of the test. Let M denote the random variable that represents the number of subjects with test diagnosis $X_i = 0, 1$ and smoking status $D_i = 0, 1$. Subsequently, let $M = M_{11}$ denote the number of true smokers using the reference test and diagnosed as smokers using the response data. Following Zhou *et al.* [36] and Broemeling [37], the conditional probability of being positively diagnosed (i.e., conditional posterior sensitivity or true positive fraction) of the self-report test given the diagnoses, $\mathbf{x}$, equals

$$SE\left(\theta_c \mid \mathbf{y}, \mathbf{x}\right) = \frac{P\left(M = M_{11} \mid \mathbf{y}, \mathbf{x}, \theta_c\right)}{P\left(M = M_{11} \mid \mathbf{y}, \mathbf{x}, \theta_c\right) + P\left(M = M_{01} \mid \mathbf{y}, \mathbf{x}, \theta_c\right)}, \tag{6}$$

where M_{01} denotes the number of smokers according to the reference test but diagnosed as non-smokers using the response data.

In the same way, the conditional probability of being negatively diagnosed (i.e., conditional posterior specificity) given the diagnoses, $\mathbf{x}$, equals

$$SP\left(\theta_c \mid \mathbf{y}, \mathbf{x}\right) = \frac{P\left(M = M_{00} \mid \mathbf{y}, \mathbf{x}, \theta_c\right)}{P\left(M = M_{00} \mid \mathbf{y}, \mathbf{x}, \theta_c\right) + P\left(M = M_{10} \mid \mathbf{y}, \mathbf{x}, \theta_c\right)}, \tag{7}$$

where M_{10} is the number of non-smokers that are positively diagnosed.

The posterior probability of random variable M follows from a sequence of Bernoulli trails. For example, given the diagnoses $\mathbf{x}$ and the outcomes of the reference test, a specific value of $M = M_{11}$ is observed, and the corresponding posterior probability can be expressed as

$$\begin{aligned} P\left(M = M_{11} \mid \mathbf{y}, \mathbf{x}, \theta_c\right) &= \int \cdots \int \prod_{j \in M_{11}} I\left(\theta_j > \theta_c\right) \prod_{h \notin M_{11}} I\left(\theta_h < \theta_c\right) p\left(\theta_1, \ldots, \theta_N \mid \mathbf{y}\right) \mathrm{d}\theta_1 \ldots \mathrm{d}\theta_N \\ &= \prod_{j \in M_{11}} \int I\left(\theta_j > \theta_c\right) p\left(\theta_j \mid \mathbf{y}_j\right) \mathrm{d}\theta_j \prod_{h \notin M_{11}} \int I\left(\theta_h < \theta_c\right) p\left(\theta_h \mid \mathbf{y}_h\right) \mathrm{d}\theta_h \\ &= \prod_{j \in M_{11}} p_{jc} \prod_{h \notin M_{11}} \left(1 - p_{hc}\right). \end{aligned} \tag{8}$$

Note that it is unlikely that the randomly selected subjects influence each other's smoking behavior. Therefore, in the first line of Equation (8), we assumed the events (i.e., smoking behavior scores are above or below the threshold) to be independent. Subsequently, we computed the probabilities of occurrence by using the marginal posterior densities of smoking behavior as specified in Equation (5).

2.3.2. Test diagnoses treated as missing data. The test diagnoses are never observable, but the posterior information about smoking behavior is used to construct the posterior distribution of the test diagnoses. Subsequently, the expected value of the posterior sensitivity (Equation (6)) and specificity (Equation (7)) can be defined, where the expectation is taken over the posterior distribution of diagnoses. This way, posterior uncertainty about the diagnoses is taken into account in the computation of posterior classification probabilities.

According to Equation (5), each random variable X_i, representing the smoking status, is Bernoulli distributed with success probability p_{ic}. The posterior probability of the diagnoses, $\mathbf{x}$, follows from independent Bernoulli trials, where the success probabilities are defined by the posterior distributions of smoking behavior; that is,

$$\begin{aligned} P\left(\mathbf{X} = \mathbf{x} \mid \mathbf{y}, \theta_c\right) &= \prod_{j:X_j=1} \int I\left(\theta_j > \theta_c\right) p\left(\theta_j \mid \mathbf{y}_j\right) \mathrm{d}\theta_j \prod_{h:X_h=0} \int I\left(\theta_h \leqslant \theta_c\right) p\left(\theta_h \mid \mathbf{y}_h\right) \mathrm{d}\theta_h \\ &= \prod_{j:X_j=1} p_{jc} \prod_{h:X_h=0} \left(1 - p_{hc}\right). \end{aligned}$$

To express the uncertainty about the diagnoses, $\mathbf{x}$, we averaged the posterior specificity and sensitivity over the posterior distribution of diagnoses. Let $\mathcal{X}$ denote the set of all possible diagnoses for the N subjects. Then we express the expected posterior sensitivity and specificity as

$$SE\left(\theta_c \mid \mathbf{y}\right) = E\left[SE\left(\theta_c \mid \mathbf{y}, \mathbf{x}\right) \mid \mathbf{y}\right] = \sum_{\mathbf{x} \in \mathcal{X}} SE\left(\theta_c \mid \mathbf{y}, \mathbf{x}\right) P\left(\mathbf{x} \mid \mathbf{y}, \theta_c\right) \tag{9}$$

and

$$SP\left(\theta_c \mid \mathbf{y}\right) = E\left[SP\left(\theta_c \mid \mathbf{y}, \mathbf{x}\right) \mid \mathbf{y}\right] = \sum_{\mathbf{x} \in \mathcal{X}} SP\left(\theta_c \mid \mathbf{y}, \mathbf{x}\right) P\left(\mathbf{x} \mid \mathbf{y}, \theta_c\right), \tag{10}$$

respectively. Finally, an expected posterior predictive value (PPV) can be defined as

$$PPV\left(\theta_c \mid \mathbf{y}\right) = \sum_{\mathbf{x} \in \mathcal{X}} \frac{P\left(M = M_{11} \mid \mathbf{y}\right)}{P\left(M = M_{11} \mid \mathbf{y}\right) + P\left(M = M_{10} \mid \mathbf{y}\right)} P\left(\mathbf{x} \mid \mathbf{y}, \theta_c\right)$$

and an expected posterior negative predictive value (NPV) as

$$NPV\left(\theta_c \mid \mathbf{y}\right) = \sum_{\mathbf{x}\in\mathcal{X}} \frac{P\left(M = M_{00} \mid \mathbf{y}\right)}{P\left(M = M_{00} \mid \mathbf{y}\right) + P\left(M = M_{01} \mid \mathbf{y}\right)} P\left(\mathbf{x} \mid \mathbf{y}, \theta_c\right).$$

The posterior expected classification probabilities can be computed as by-products of the MCMC algorithm. In a first step, smoking diagnoses are sampled given a threshold θ_c, according to Equation (5), using the posterior distribution of smoking behavior. In iteration m, the sample $\mathbf{x} = \mathbf{x}^{(m)}$ is used to compute the conditional posterior sensitivity and conditional posterior specificity probabilities, as defined in Equations (6) and (7). The posterior expected specificity and expected sensitivity, Equations (9) and (10), are estimated by the average values over MCMC iterations.

3. Results

We conducted the 37-day survey in cooperation with 10 pulmonologists. We asked lung patients older than 16 years to voluntarily participate in the survey on smoking. None of the randomly selected patients refused to cooperate. We assessed a total of 305 patients, and 198 patients completed the test using the RRT. The RR group was significantly larger than the DQ group to account for the generated forced responses. For the items with two response categories, the developed random device requested an honest answer with probability 0.778 and simulated each forced response (yes and no) with probability 0.5. For the items with three response categories, we requested an honest response with probability 0.611, and the forced category response probabilities were 0.25 for the first and third categories and 0.5 for the second category. We used CO to detect smokers, and we took a cut-off value of 6 parts per million (ppm) as the cutoff between smokers ($D = 1$) and non-smokers ($D = 0$).

From the classical test procedure, for the 12-item direct-questioning data, Cronbach's alpha equals 0.78, which indicates that the scale has a good reliability. Cronbach's alpha cannot be directly computed from the RR data because of the forced responses. Following the correction method of Himmelfarb [38], Cronbach's alpha equals 0.86. It can be concluded that the items correlate slightly better under the RR questioning technique.

Table I. Smoking behavior study: Estimated model parameters for RR and DQ data.

	Item	Mean	SD	95% CI	
Threshold parameters					
b_1	1	0.693	0.124	0.444	0.935
b_2	2	0.645	0.119	0.407	0.873
b_3	3	1.094	0.138	0.841	1.366
b_4	4	−0.321	0.114	−0.556	−0.094
b_5	5	2.143	0.206	1.768	2.563
b_6	6	0.513	0.121	0.284	0.762
b_7	7	1.328	0.154	1.046	1.634
b_8	8	0.299	0.120	0.062	0.543
b_9	9	−0.132	0.117	−0.366	0.102
κ_1	10	−1.498	0.151	−1.803	−1.203
κ_2	10	−0.749	0.131	−1.008	−0.497
κ_1	11	0.656	0.132	0.402	0.924
κ_2	11	1.557	0.164	1.245	1.888
κ_1	12	0.625	0.131	0.359	0.884
κ_2	12	0.817	0.133	0.555	1.077
Population parameters					
σ^2_θ		0.953	0.133	0.721	1.239
σ^2_b		0.836	0.498	0.319	2.066
μ_b		0.691	0.312	0.073	1.310
π_{01}		0.037	0.022	0.004	0.086
π_{02}		0.017	0.013	0.001	0.049

Note: DQ, directly questioned; RR, randomized response.

We fitted the mixture randomized item-response model using WinBUGS [39], with a burn-in period of 5000 followed by 10,000 iterations. The computer code is given in Appendix C.

Table I represents the model parameter estimates. The estimated item characteristics span a wide range, which is useful for assessing accurately different smoking behaviors. For the binary items, the item population distribution has a mean of 0.691 and a variance of 0.836, which indicates the relatively large variation in item difficulties. The behavior population distribution has a mean of 0, to identify the latent scale, and a variance of 0.953. Participants endorsing the items with high values of threshold parameters score higher on the latent scale. The low estimated thresholds of item 10 reveal that most subjects confirmed to have smoked at least a few years, where most of them (around 80%) smoked more than 15 years. Most of the patients are aware that smoking is an unhealthy habit, which follows from the scores on item 5.

We implemented mixture components (compliant and non-compliant class) for the binary and polytomous items to control for positive self-report behavior of subjects not following the RR instructions. From the rating scale specific non-compliance probabilities follows that around 3.7% (π_{01}) and 1.7% (π_{02}) of the patients were detected as non-complying given the binary and polytomous response patterns, respectively.

3.1. *RRT validation*

We randomly assigned participants to the RR group or the DQ group. Therefore, it was expected that mean behavior differences were only attributable to the questioning technique. Furthermore, we expected scores from respondents in the RR condition to correlate more strongly with the (physical) CO measurement than those in the DQ condition because of the induced guarantee of confidentiality inherent to the RRT.

We defined two models to explore the effects of the treatment condition (RR), the CO measurement, and their interaction on subject's smoking behavior. Model 1 only contains the main effects. Model 2 also contains the interaction effect such that the latent regression component matches the one displayed in Equation (4). We adjusted the WinBUGS program in Appendix C to estimate the parameters of both models. The regression parameter estimates of both models are reported in Table II.

Model 1 estimates show that subjects in the RR condition scored significantly higher, on average around 0.39, than those in the DQ condition. Subjects diagnosed as smokers ($CO = 1$) scored on average 1.62 higher than those diagnosed as non-smokers. This strong relationship between the estimated smoking behavior scores and the CO measurements indicates that the response data contain information about the smoking status (smokers and non-smokers) of subjects. The intercept represents the average smoking behavior of diagnosed non-smokers ($CO = 0$) in the DQ group, which was scaled around 0 to identify the mean of the scale.

We used the estimated interaction effect in model 2 to evaluate the differential effect of questioning techniques for subjects diagnosed as smokers ($CO = 1$) and non-smokers ($CO = 0$) on the smoking behavior scores. It follows that in the RR condition, smokers ($CO = 1$) scored on average 1.37 plus 0.50 higher than the non-smokers. In the DQ condition, smokers ($CO = 1$) scored on average 1.37 higher than

Table II. Smoking behavior study: Influence of RRT and CO measurement.

	Model 1			Model 2		
	Mean	SD	CI	Mean	SD	CI
Parameter						
β_0 (Intercept)	0.00	0.10	−0.22, 0.19	0.00	0.08	−0.20, 0.19
β_1 (RR)	0.39	0.09	0.20, 0.57	0.22	0.11	0.00, 0.44
β_2 (CO)	1.62	0.10	1.41, 1.84	1.37	0.14	1.09, 1.66
β_3 (RR × CO)				0.50	0.20	0.11, 0.90
Population parameters						
σ^2_θ	0.28	0.05	0.19, 0.39	0.28	0.05	0.18, 0.40
π_{01}	0.04	0.02	0.00, 0.09	0.03	0.02	0.00, 0.08
π_{02}	0.01	0.01	0.00, 0.04	0.01	0.01	0.00, 0.04

Note: RRT, randomized response technique; CI, 95% credible interval.

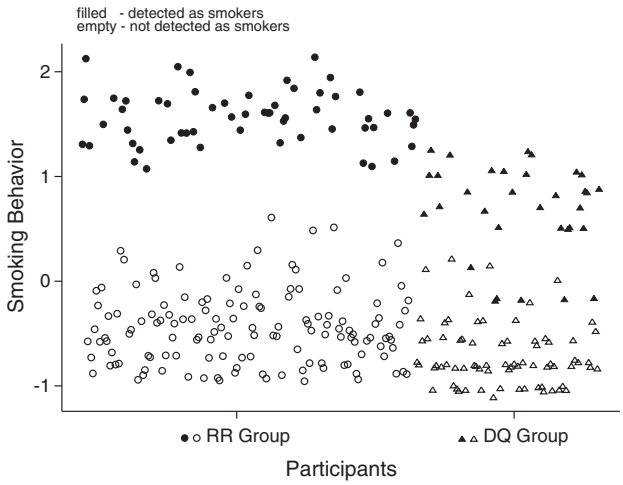

Figure 1. Estimated smoking behavior scores of smokers and non-smokers, diagnosed with the CO measurement, in the DQ and RR conditions.

the non-smokers. For the non-smokers ($CO = 0$), this difference in scores over questioning techniques was much lower, where non-smokers ($CO = 0$) in the RR condition scored 0.22 higher than non-smokers in the DQ condition.

The items were not very sensitive for the non-smokers, and as a result, the difference in smoking behavior scores between questioning techniques was relatively small. We based the smoking status on the CO measure. When a smoker abstains for a period of more than 12 h, this smoker could be erroneously tested as a non-smoker. Therefore, for diagnosed non-smokers ($CO = 0$), the difference in scores between questioning techniques might be positively biased because smokers were incorrectly classified as non-smokers. However, it was unlikely to classify non-smokers as smokers, which diminished the possibility that such misclassifications led to a downwardly biased interaction effect.

The more interesting case concerns the smokers in both conditions. The estimated interaction effect is around two times higher than the main effect of RR, which means that exactly smokers scored significantly higher in the RR condition. In the RR condition, the smokers scored significantly higher than the non-smokers, which corresponded to the CO-measured smoking status, and as expected, smokers experienced the items as more sensitive than non-smokers. This result validates the RRT.

In Figure 1, we plotted the Expected A Posteriori (EAP) scores for the smokers ($CO = 1$) and non-smokers in the DQ and RR groups, where the mean score equals 0. It follows that the estimated scores of the non-smokers in the DQ and RR groups do not differ much, but they differ for the smokers. Furthermore, the estimated scores differ more between smokers and non-smokers in the RR group than in the DQ group. This illustrates that the RRT improves the accuracy of the self-reports for those that experience the items as sensitive.

Pulmonologists have different specializations and are working on a fixed-shift basis, which could cause a clustering of patients in days. Furthermore, hospital visits of patients are often synchronized such that patients with comparable background characteristics are more likely to visit the hospital on the same day. When accounting for the additional clustering of patients in days through a random day effect, the estimated small clustering effect hardly influenced the regression effects.

We collected different background information to explain variation in smoking behavior. Therefore, we extended the regression component in Equation (4) with explanatory variables: gender, education, age, and type of disease. It was not possible to simultaneously investigate all main effects and (higher level) interactions because of the relatively small data set. We did not detect any significant differences between men and women in scores conditional on the mean structure in Equation (4). We also did not find a significant effect of education.

We considered three age groups: 17–40, 41–60, and above 60 years, where the last group was the baseline. In Table III, the parameter estimates of the main and interaction effects with RR under model 3 are given. It can be seen that the middle group scored lower, but the effect is not significantly different from 0. The positive significant interaction effect of the middle age group with RRT reveals that subjects

Table III. Smoking behavior study: Influence of age and type of disease.

	Model 3		Model 4	
	Mean	SD	Mean	SD
Parameter				
Intercept	0.00	0.10	0.00	0.09
RR	0.19	0.15	0.13	0.14
CO status	1.43	0.15	1.34	0.15
RR × CO status	0.42	0.20	0.48	0.20
Age 17–40 years	0.03	0.17		
Age 41–60 years	−0.14	0.12		
Age 17–40 years × RR	−0.27	0.23		
Age 41–60 years × RR	0.33	0.16		
Asthma			−0.30	0.18
Lung cancer			−0.17	0.40
Bronchitis			−0.34	0.38
Other			−0.43	0.17
Asthma × RR			−0.13	0.24
Lung cancer × RR			0.40	0.47
Bronchitis × RR			0.08	0.46
Other × RR			0.45	0.26
σ^2	0.28	0.06	0.27	0.05
π_{01}	0.03	0.02	0.03	0.02
π_{02}	0.01	0.01	0.01	0.01

Note: RR represents randomized response condition.

in this age group scored significantly higher under protection of the RRT. The RRT proves to be more powerful for this age group.

The different diseases were categorized as follows: asthma, lung cancer, bronchitis, and other. It follows that patients classified in group other (group effect −0.43) scored significantly lower under direct questioning, according to the estimates of model 4. However, they scored substantially higher in the RR group. Although not significant, this pattern is apparent for all groups except the disease group asthma.

3.2. Bayesian diagnostic evaluation of randomized response testing

To compute basic measures of diagnostic test accuracy of the questionnaire under the RR design, we excluded the CO measure from the measurement model. We computed the expected number of diagnosed smokers in the sample, given the response data but independent of the observed CO measures. Subsequently, we used the CO measures to validate the accuracy of the model predictions on the basis of the test data.

According to Equation (5), we computed the posterior probability of smoking per subject given the response data for threshold values ranging from 0 to 2. This interval covers the posterior smoking behavior estimates above the mean. Subsequently, for each subject and threshold value, posterior predictive diagnoses were simulated under the model. We compared each posterior predictive sample of diagnoses, $\mathbf{x}$, on the basis of the response data, with the true diagnoses on the basis of the CO measure.

In Figure 2, the estimated expected posterior sensitivity (true positive fraction) is given per threshold value together with 95% highest posterior density (HPD) intervals. It can be seen that for small threshold values, around the latent population mean, the posterior probability of diagnosing smokers correctly on basis of the response data is around 85%, and for lower threshold values (not plotted), this probability goes to 1. When increasing the higher threshold value, the true positive fraction decreases and the uncertainty increases, which follows from the increase of the HPD interval. The posterior latent mean of subjects in the RR group detected as smokers equals 1.31. When the threshold value equals this posterior latent mean ($\theta_c = 1.3$), the sensitivity is around 0.43. Each sensitivity value is based on smoking diagnoses simulated from the smoking behavior posterior distributions. The additional posterior uncertainty of each subject-specific smoking behavior causes the sensitivity to decrease more slowly towards 0 as the threshold value increases.

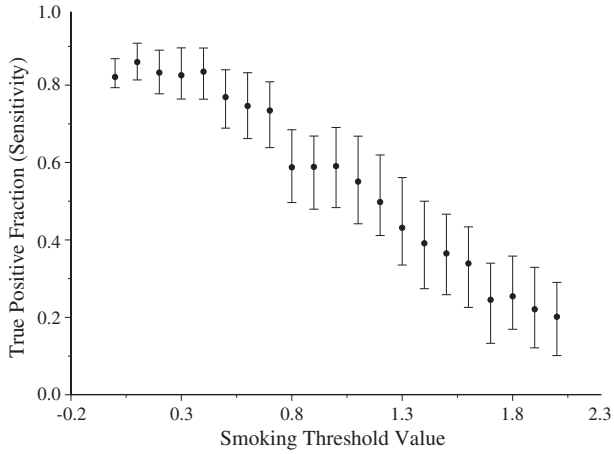

Figure 2. The posterior sensitivity as a function of threshold values from 0 to 2 using response data from subjects in the RR group.

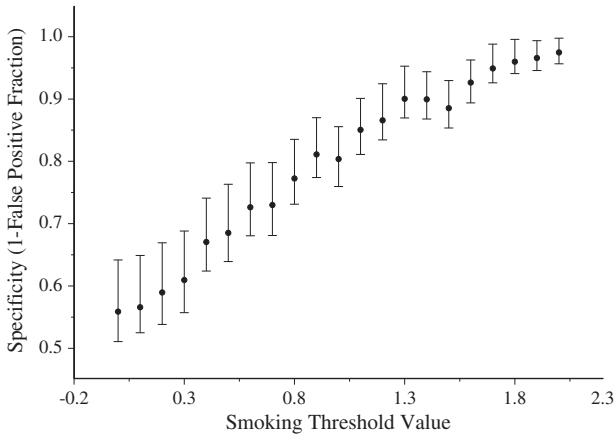

Figure 3. The posterior specificity as a function of threshold values from 0 to 2 using response data from subjects in the RR group.

In Figure 3, the estimated posterior specificity (one minus false positive fraction) and 95% HPD interval is given per threshold value using the RR data. For high threshold values, the posterior probability of smoking goes to 0 because all subjects are classified as non-smokers. The HPD intervals are relatively small as the posterior smoking behavior distributions support the decision of classifying subjects as non-smokers for such high threshold values. The probability of correctly classifying non-smokers decreases when the threshold value decreases, where the uncertainty increases. When the threshold value equals 1.3, the posterior specificity equals 0.90. When the threshold value equals 0.70, both the sensitivity and specificity are around 0.73. The posterior smoking behavior estimates are subject to considerable uncertainty because of the limited number of items and noise introduced by the RR mechanism. Together with the limited number of subjects in the study, the optimal posterior classification probabilities (specificity and sensitivity) are around 0.73.

For $\theta_c = 0.7$, the posterior mean PPV equals 0.76 for the RR data and 0.60 for the DQ data. In the same way, the posterior NPV equals 0.89 and 0.84 for the RR and DQ groups, respectively. For a perfect test, both probabilities are equal to 1, but the aforementioned different types of uncertainty reduce both predictive probability estimates. Furthermore, the posterior distributions of the PPV and NPV are skewed to the left because of the natural upper bound of 1. The sampling-based algorithm renders posterior means, which are most likely conservative underestimates of the true values.

A comparison between the DQ and RR group of posterior classification probabilities is complicated because of scale differences, which makes it problematic to define one common set of threshold values over groups. Furthermore, the DQ group shows serious underreporting, which increases the probability of classifying non-smokers correctly. The consistent low scores in the DQ group significantly

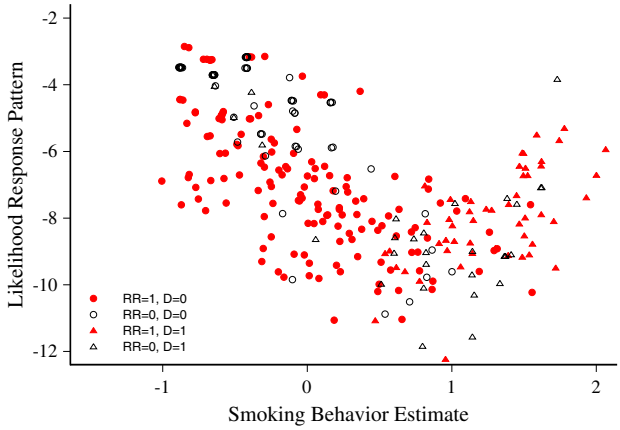

Figure 4. For each subject, the estimated smoking behavior against the log-likelihood of the response pattern.

improves the test accuracy when it also concerns non-smokers, which are 71% of the respondents in the DQ group. From that perspective, the test accuracy only increases using the RRT when it concerns sensitive information.

In Figure 4, the estimated smoking behavior scores of smokers ($D = 1$) and non-smokers ($D = 0$) in the DQ and RR conditions are plotted against the log-likelihood of the corresponding response patterns. For the non-smokers ($D = 0$), it can be seen that a relatively large group of non-smokers in the RR condition have relatively low log-likelihood values. This indicates that on average the model fits the non-smokers' response patterns in the DQ condition better than those in the RR condition. However, the response patterns of smokers ($D = 1$) in the RR condition have higher log-likelihood values than those of smokers in the DQ condition. On average, the model fits the response patterns of smokers in the RR condition better than those of smokers in the DQ condition. Smokers' response patterns in the RR condition are on average more consistent than those in the DQ condition, and as a conclusion, the RRT improves the accuracy of smoker's self-reports.

4. Discussion and conclusions

In the present sensitive survey study, we assessed smoking behavior of outpatients of a Dutch pulmonary department using different questioning techniques. It was shown that the randomized-response technique led to more accurate responses in comparison with the direct-questioning technique. Therefore, we used CO measurements as a gold standard to evaluate test diagnoses based on (randomized) item response data. Participants detected as smokers scored significantly higher in the RR condition compared with smokers in the DQ condition. Such an apparent difference in scoring was not detected for the non-smokers, although non-smokers scored slightly higher in the RR condition. Non-smokers did not perceive the smoking questionnaire as sensitive, and the questioning technique hardly influenced their test results. The randomized-response technique positively influenced the quality of smokers' response data according to the gold standard, which validates the multi-item randomized-response technique.

In experimental studies using RRT, a treatment–control design is often used for validation of the results, where the randomly selected members of the control and the randomized-response group are interviewed via direct or RR questioning, respectively. Without knowing the truth, a difference between the mean response of the DQ and the RR group cannot be interpreted as a validation of the item randomized-response technique. Umesh and Peterson [40] commented that a legitimate validation of RR estimates will rest on true response values, preferably at the individual level.

Although not for the item randomized-response technique, a few studies did use true individual information about the sensitive behavior to validate the (univariate) randomized-response technique. Van der Heijden *et al.* [24] reported about respondents caught for welfare or unemployment benefit fraud and investigated the percentage of underreporting without having a control group. Akers *et al.* [1] used a physiological measure, salivary thiocyanate, to identify a respondent's smoking behavior. They investigated smoking behavior among adolescents and found approximately similar results with

both the randomized-response and direct-questioning methods. They concluded that the participants might have experienced smoking as non-sensitive behavior. However, the participants were informed that saliva samples were being collected after the survey, which was to convince the subjects that their self-reports could be verified. As a result, this feature of the study design stimulated truthful reporting in both RR and DQ groups. This so-called bogus pipeline [41] probably diminished the effect of the RRT.

The developed mixture randomized item-response model enabled the simultaneous analysis of mixed item response data (binary and ordinal) that were obtained using the RR or DQ technique. We pursued a mixture modeling approach to account for subjects that ignore the RR instruction and consistently scored in the least incriminating response category. Therefore, we defined a non-compliance group as a separate latent class for the ordinal and binary response data.

The subject-specific posterior probability that the (continuous) smoking behavior is greater than a specified threshold value defined the posterior probability of smoking. We used this posterior probability distribution to simulate smoking diagnoses and to evaluate the diagnostic accuracy of the self-report smoking questionnaire in the RR condition. This innovative latent variable approach enabled the computation of the posterior sensitivity and specificity given self-reported RR data and CO measurements. Further research in this area is needed to investigate the influence of different randomized-response techniques, test length, and item sensitivity on the test accuracy. The presented Bayesian latent variable framework for evaluating the test accuracy can also be extended to ordinal diagnostic measurements.

Appendix A. Forced randomized response: sequence of Bernoulli trials

The response model given in Equation (2) is explained in detail, and the probability distribution of the randomized item response variable is derived. It is shown that the forced RR data collection technique consists of a sequence of independent Bernoulli trials to mask individual responses.

First, a Bernoulli trial defines the RR experiment for each item k answered by subject i. This trial is described by random variable Z_{ik}, which has a Bernoulli distribution and can take the value 0 (i.e., forced response requested), with probability $1-\phi_1$ or 1 (i.e., honest response requested), with probability ϕ_1. Second, given the outcome of Z_{ik}, a second independent Bernoulli trial is performed. For $Z_{ik}=1$, an honest response is requested, and the Bernoulli trial is described by random variable $\tilde{Y}_{ik}$. For $Z_{ik}=0$, a forced response is requested, and the Bernoulli trial is described by random variable F_{ik}, which takes the value 1 with probability ϕ_2 and 0 with probability $1-\phi_2$. Third, a random variable G_{ik} is defined, which takes the value 1, with probability π_i^b, when subject i ignores the RR instruction and responds in the least self-incriminating way, and 0 otherwise with probability $1-\pi_i^b$.

The sequence of independent Bernoulli trials leads to the definition of an RR variable, denoted as Y_{ik}. If subject i follows the RR instructions ($G_{ik}=0$), an honest response of 1 when requested ($Z_{ik}=1$), or a forced response of 1 when requested ($Z_{ik}=0$), will lead to an RR of 1; that is,

$$I(Y_{ik}=1)=I(G_{ik}=0)\left[I(Z_{ik}=1)I(\tilde{Y}_{ik}=1)+I(Z_{ik}=0)I(F_{ik}=1)\right]. \tag{11}$$

The random variable Y_{ik} describes the outcome of the sequence of Bernoulli trials and is Bernoulli distributed with success probability

$$\begin{aligned}P(Y_{ik}=1)&=P(G_{ik}=0)\left(P(Z_{ik}=1)P(\tilde{Y}_{ik}=1)+P(Z_{ik}=0)P(F_{ik}=1)\right)\\&=\left(1-\pi_i^b\right)\left(\phi_1 P(\tilde{Y}_{ik}=1)+(1-\phi_1)\phi_2\right).\end{aligned}$$

Let the honest response to item k of subject i be described by the one-parameter item-response model with success probability $\Phi(\theta_i-b_k)$. Then the success probability of the event $Y_{ik}=1$ is given by

$$P(Y_{ik}=1\mid\theta_i,b_k)=\left(1-\pi_i^b\right)\left(\phi_1 P(\tilde{Y}_{ik}=1\mid\theta_i,b_k)+(1-\phi_1)\phi_2\right)$$
$$=\left(1-\pi_i^b\right)\left(\phi_1\Phi(\theta_i-b_k)+(1-\phi_1)\phi_2\right),$$

which resembles the success probability of subject i in the RR condition of Equation (2). In the same way, the probability of an RR of 0 can be derived, which is given by,

$$\begin{aligned}P(Y_{ik}=0\mid\theta_i,b_k)&=P(G_{ik}=1)+P(G_{ik}=0)\left(P(Z_{ik}=1)P(\tilde{Y}_{ik}=0\mid\theta_i,b_k)\right.\\&\quad\left.+P(Z_{ik}=0)P(F_{ik}=0)\right)\\&=\pi_i^b+\left(1-\pi_i^b\right)\left(\phi_1(1-\Phi(\theta_i-b_k))+(1-\phi_1)(1-\phi_2)\right)\\&=\pi_i^b+\left(1-\pi_i^b\right)\left(1-(\phi_1\Phi(\theta_i-b_k)+(1-\phi_1)\phi_2)\right),\end{aligned}$$

which resembles the failure probability of subject i in the RR condition of Equation (2).

Appendix B. Smoking behavior scale questionnaire

The questionnaire comprises nine dichotomous items followed by three polytomous items. In the RR condition, for each item, a spinner was spun, and wherever the arrow landed determined whether the item was to be answered honestly or dictated the answer choice to be recorded by the participant. For participants with a condition called color blindness (of any type), a modified digitized spinner disc was available.

1. Do you smoke?
2. Do you feel like smoking when having a cup of coffee or a glass of beer?
3. Do you feel like smoking when somebody else is smoking next to you?
4. Does it disturb you when somebody else is smoking in the same room? (Score reversed)
5. Do you think that smoking is an unhealthy habit? (Score reversed)
6. Do you think that smoking is disgusting? (Score reversed)
7. Do you enjoy the smell of somebody else's cigarette?
8. Do you support the governmental policy concerning the indoor smoking ban? (Score reversed)
9. Are your guests allowed to smoke at your house?
10. How many years have you been smoking/had been smoking? (response categories: never, 15 or less, more than 15)
11. How may cigarettes are you smoking per day? (none, 10 or less, more than 10)
12. How many days per week are you smoking? (none, 4 or less, more than 4)

Appendix C. WinBUGS code: mixture randomized item-response model

The code of the mixture randomized item-response model for RR and DQ data is given for N persons and K dichotomous and Kp polytomous items. The randomizing device parameters are p1 and p2 for dichotomous response data, and pp1 and pp2(c) for polytomous response data. The indicator RR[i] equals 0 (1) when participant i belongs to the DQ group (RR group).

```
model{
   for(i in 1:N){ #subjects
     for(k in 1:K){ #dichotomous items
         Q[i,k] <-phi(theta[i] - b[k])
         p[i,k,1]<-(1-RR[i])*(1-Q[i,k])+ RR[i]*(pi1[i]+(1-pi1[i])*(1-(p1*Q[i,k]+(1-p1)*p2)))
         p[i,k,2]<-(1-RR[i])*Q[i,k] + RR[i]*(1-pi1[i])*(p1*Q[i,k]+(1-p1)*p2)
         Y[i,k] ~ dcat(p[i,k,1:2])
     }
     for(kk in 1:Kp){#polytomous items
        for (c in 1:2){#response categories
            Qp[i,kk,c] <- phi(bp[kk,c]-theta[i])
        }
        pp[i,kk,1] <- (1-RR[i])*Qp[i,kk,1]+RR[i]*(pi2[i]+(1-pi2[i])*(pp1*Qp[i,kk,1]+(1-pp1)*pp2[1]))
        pp[i,kk,2] <- (1-RR[i])*(Qp[i,kk,2]-Qp[i,kk,1])+RR[i]*(1-pi2[i])*(pp1*(Qp[i,kk,2]-Qp[i,kk,1])+
            (1-pp1)*pp2[2])
        pp[i,kk,3] <- (1-RR[i])*(1-Qp[i,kk,2])+RR[i]*(1-pi2[i])*(pp1*(1-Qp[i,kk,2])+(1-pp1)*pp2[3])
        Yp[i,kk] ~ dcat(pp[i,kk,1:3])
      }
      pi1[i] ~ dbern(pi01) #mixture model dichotomous data
      pi2[i] ~ dbern(pi02) #mixture model polytomous data
      theta[i] ~ dnorm(mutheta[i],sigmathetaP)
      mutheta[i] <- beta[1] + beta[2]*RR[i] + beta[3]*CO[i]
     }
     #prior distributions
     for(k in 1:K){
        b[k]~dnorm(mub,sigmabP)
     }
     for(kk in 1:Kp){
        bp[kk,1]<-mu[kk]
        bp[kk,2]<-bp[kk,1] + exp(mu2[kk])
        mu[kk]~dnorm(-1,0.1)
        mu2[kk]~dnorm(0,1)
     }
     pi01~dbeta(1,1)
     pi02~dbeta(1,1)
     beta[1] ~ dnorm(0,100)   #identification mean scale
     beta[2] ~ dnorm(0,1.0E-1)
     beta[3] ~ dnorm(0,1.0E-1)
     sigmathetaP ~ dgamma(1,1)
     sigmatheta <- 1/sigmathetaP
     mub ~ dnorm(0,1.0E-2)
     sigmabP ~ dgamma(1,1)
     sigmab <- 1/sigmabP
  }
```

References

1. Akers RL, Massey J, Clarke W, Lauer RM. Are self-reports of adolescent deviance valid? Biochemical measures, randomized response and the bogus pipeline in smoking behavior. *Social Forces* 1983; **62**:234–251. DOI: 10.1093/sf/62.1.234.
2. Monninkhof EM, van der Valk PD, van der Palen J, Mulder H, Pieterse M, van Herwaarden CL, Zielhuisd G. The effect of a minimal contact smoking cessation programme in out-patients with chronic obstructive pulmonary disease: a pre–post-test study. *Patient Education and Counseling* 2004; **52**:231–236. DOI: 10.1016/S0738-3991(03)00096-X.
3. Attebring M, Herlitz J, Berndt A-K, Karlsson T, Hjalmarson A. Are patients truthful about their smoking habits? A validation of self-report about smoking cessation with biochemical markers of smoking activity amongst patients with ischaemic heart disease. *Journal of Internal Medicine* 2001; **249**:145–151. DOI: 10.1046/j.1365-2796.2001.00770.x.
4. Tourangeau R, Yan T. Sensitive questions in surveys. *Psychological Bulletin* 2007; **133**:859–883. DOI: 10.1037/0033-2909.133.5.859.
5. Daly RJ, Blann AD. Self-reported smoking in vascular disease: the need for biochemical confirmation. *British Journal of Biomedical Science* 1996; **53**:204–208.
6. Hill P, Haley NJ, Wynder EL. Cigarette smoking: carboxyhemoglobin, plasma nicotine, cotinine and thiocyanate vs self-reported smoking data and cardiovascular disease. *Journal of Chronic Diseases* 1983; **6**:439–449. DOI: 10.1016/0021-9681(83)90136-4.
7. Jarvis MJ, Tunstall-Pedoe H, Feyerabend C, Vesey C, Saloojee Y. Comparison of tests used to distinguish smokers from nonsmokers. *American Journal of Public Health* 1987; **11**:1435–1438. DOI: 10.2105/AJPH.77.11.1435.
8. Warner SL. Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* 1965; **60**:63–69. DOI: 10.1080/01621459.1965.10480775.
9. Soeken KL. Randomized response research in health research. *Evaluation & the Health Professions* 1987; **10**:58–66. DOI: 10.1177/016327878701000105.

10. Rittenhouse BE. A novel compliance assessment technique—the randomized response interview. *International Journal of Technology Assessment in Health Care* 1996; **12**:498–510. DOI: 10.1017/S0266462300009843.
11. Rittenhouse BE. Respondent-specific information from the randomized response interview: compliance assessment. *Journal of Clinical Epidemiology* 1996; **49**:545–549. DOI: 10.1016/0895-4356(96)00005-4.
12. Williams BL, Suen HK, Baffi CR. A controlled randomized response technique. *Evaluation & the Health Professions* 1993; **16**:225–245. DOI: 10.1177/016327879301600207.
13. Cross P, Edwards-Jones G, Omed H, Williams AP. Use of a randomized response technique to obtain sensitive information on animal disease prevalence. *Preventive Veterinary Medicine* 2010; **96**:252–262. DOI: 10.1016/j.prevetmed.2010.05.012.
14. Ostapczuk M, Musch J, Moshagen M. Improving self-report measures of medication non-adherence using a cheating detection extension of the randomized-response-technique. *Statistical Methods in Medical Research* 2011; **20**:489–503. DOI: 10.1177/0962280210372843.
15. Fox J-P. Randomized item response theory models. *Journal of Educational and Behavioral Statistics* 2005; **30**:1–24. DOI: 10.3102/10769986030002189.
16. De Jong MG, Pieters R, Fox J-P. Reducing social desirability bias via item randomized response: an application to measure underreported desires. *Journal of Marketing Research* 2010; **47**:14–27. DOI: 10.1509/jmkr.47.1.14.
17. De Jong MG, Pieters R, Stremersch S. Analysis of sensitive questions across cultures: an application of multigroup item randomized response theory to sexual attitudes and behavior. *Journal of Personality and Social Psychology* 2012; **103**:543–564. DOI: 10.1037/a0029394.
18. Fox J-P, Meijer RR. Using item response theory to obtain individual information from randomized response data: an application using cheating data. *Journal of Applied Psychological Measurement* 2008; **32**:595–610. DOI: 10.1177/0146621607312277.
19. Fox J-P, Wyrick C. A mixed effects randomized item response model. *Journal of Educational and Behavioral Statistics* 2008; **33**:389–415. DOI: 10.3102/1076998607306451.
20. Tracy PE, Fox JA. The validity of randomized response for sensitive measurements. *American Sociological Review* 1981; **46**:187–199.
21. Scheers NJ, Dayton C. Covariate randomized response model. *Journal of the American Statistical Association* 1988; **83**:969–974. DOI: 10.1080/01621459.1988.10478686.
22. Lensvelt-Mulders G. JLM, Hox JJ, Van der Heijden PGM, Maas C. Meta-analysis of randomized response research: 35 years of validation. *Sociological Methods & Research* 2005; **33**:319–348. DOI: 10.1177/0049124104268664.
23. Lensvelt-Mulders GJLM, Hox JJ, van der Heijden PGM. How to improve the efficiency of randomized response designs. *Quality and Quantity* 2005; **39**:253–265. DOI: 10.1007/s11135-004-0432-3.
24. van der Heijden PGM, van Gils G, Bouts J, Hox JJ. A comparison of randomized response, computer-assisted self-interview, and face-to-face direct questioning: eliciting sensitive information in the context of welfare and unemployment benefit. *Sociological Methods & Research* 2000; **28**:505–537. DOI: 10.1177/0049124100028004005.
25. Lord FM, Novick MR. *Statistical Theories of Mental Test Scores*. Addison-Wesley: Oxford, England, 1968.
26. Clark SJ, Desharnais RA. Honest answers to embarrassing questions: detecting cheating in the randomized response model. *Psychological Methods* 1998; **3**:160–168. DOI: 10.1037/1082-989X.3.2.160.
27. Böckenholt U, van der Heijden PGM. Item randomized-response models for measuring non-compliance: risk-return perceptions, social influences, and self-protective responses. *Psychometrika* 2007; **72**:245–262. DOI: 10.1007/s11336-005-1495-y.
28. Fox J-P. *Bayesian Item Response Modeling: Theory and Applications*. Springer: New York, 2010. DOI:10.1007/978-1-4419-0742-4.
29. Shadel W, Shiffman S. Assessment of smoking behavior. In *Assessment of Addictive Behaviors*, Donovan DM, Marlatt GA (eds), 2nd edn, chap. 4. Guilford: New York, 2005; 113–154.
30. Heatherton TF, Kozlowski LT, Frecker RC, Fagerström K.-O. The Fagerström test for nicotine dependence: a revision of the Fagerström Tolerance Questionnaire. *British Journal of Addiction* 1991; **86**:1119–1127.
31. Wetter DW, Kenford SL, Welsch SK, Smith SS, Fouladi RT, Fiore MC, Baker TB. Prevalence and predictors of transitions in smoking behavior among college students. *Health Psychobiology* 2004; **23**:168–177. DOI: 10.1037/0278-6133.23.2.168.
32. Middleton ET, Morice AH. Breath carbon monoxide as an indication of smoking habit. *Chest* 2000; **117**:758–763. DOI: 10.1378/chest.117.3.758.
33. Tate JC, Schmitz JM. A proposed revision of the Fagerström Tolerance Questionnaire. *Addictive Behaviors* 1993; **18**:135–143. DOI: 10.1016/0306-4603(93)90043-9.
34. Garrett ES, Eaton WW, Zeger S. Methods for evaluating the performance of diagnostic tests in the absence of a gold standard: a latent class model approach. *Statistics in Medicine* 2002; **21**:1289–1307. DOI: 10.1002/sim.1105.
35. Formann KA, Kohlmann T. Latent class analysis in medical research. *Statistical Methods in Medical Research* 1996; **5**:179–211. DOI: 10.1177/096228029600500205.
36. Zhou XH, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine*, 2nd edn. Wiley: New York, 2002.
37. Broemeling LD. *Bayesian Biostatistics and Diagnostic Medicine*. Chapman and Hall: Boca Raton, FL, 2007.
38. Himmelfarb S. The multi-item randomized response technique. *Sociological Methods and Research* 2008; **36**:495–514. DOI: 10.1177/0049124107313900.
39. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 2000; **10**:325–337. DOI: 10.1023/A:1008929526011.
40. Umesh UN, Peterson RA. A critical-evaluation of the randomized-response method—applications, validation, and research agenda. *Sociological Methods & Research* 1991; **20**:104–138. DOI: 10.1177/0049124191020001004.
41. Jones E, Sigall H. The bogus pipeline: a new paradigm for measuring affect and attitude. *Psychological Bulletin* 1971; **76**:349–364. DOI: 10.1037/h0031617.