1

Bayesian Randomized Item Response Theory Models for Sensitive Measurements

CONTENTS

1.1 Introduction

Research on behavior and attitudes typically relies on self-reports, especially when the infrequency of behavior and research costs make it hardly impractical not to do so. However, many studies have shown that self-reports can be highly unreliable and actually serve as a fallible source of data.

Results from self-reports are often influenced by such factors as the question order, wording, or response format, even when they contain simple behavioral questions. In general, the psychology of asking questions has received considerable attention in the literature (e.g., Sudman, Bradburn, & Schwarz, 1996; Tourangeau, Rips, & Rasinski, 2000), and a growing body of research has provided sound methods of question development that do improve the quality of self-report data.

Finally, the quality of self-report data depends on the respondents' willingness to cooperate and give honest answers. Especially for sensitive topics, it is known that people tend to report in a socially desirable way; that is, in the direction of the re-

searcher's expectations and/or what reflects positively on their behavior. Thus, sensitivity of the questions easily leads to misreporting (i.e. under- or overreporting), even when anonymity and confidentiality of the responses is guaranteed. This observation is supported by considerable empirical evidence. For instance, survey respondents underreported socially undesirable behavior such use of illicit drugs (Anglin, Hser, & Chou, 1993), the number of sex partners (Tourangeau & Smith, 1996), desires for adult entertainment (De Jong, Pieters, & Fox, 2010), welfare fraud (van der Heijden, van Gils, Bouts, & Hox, 2000), and alcohol abuse and related problems (Fox & Wyrick, 2008).

In order to overcome respondents' tendencies to report inaccurately or even to refuse to provide any response at all, strategies have been developed to deal with them. Typically, anonymity of the respondents is guaranteed and explicit assurances are given that each of their answers will remain completely confidential. Besides, questions are often phrased such that tendencies to provide socially desirable answers are diminished. Furthermore, respondents are motivated to provide accurate answers by stressing the importance of the research study.

Other ways of avoiding response tendencies to report inaccurately are based on innovative data collection methods that make it impossible to infer any identifying information from the response data. A general class of such methods for sensitive surveys is based on the randomized response technique (RRT) (Fox & Tracy, 1986), which involves the use of a randomizing device to mask individual responses. RRT has originated from Warner (1965), who developed a randomized response (RR) data collection procedure, where respondents are confronted with two mutually exclusive questions, for instance, "I belong to Group A," and "I do not belong to Group A." A choice is made between the two statements using a randomizing device (e.g., tossing of a die or use of a spinner). The randomization is performed by the respondent and the outcome is not revealed to the interviewer. The respondent then answers the question selected by the randomizing device. The interviewer only knows the response, not the question.

Because of this setup, the RR technique encourages greater cooperation from respondents and reduces socially desirable response behavior. The properties of the randomizing device are known, which still allows for population estimates of the sensitive behavior, for instance, proportions of the population engaging in a particular kind of behavior or, more generally, membership of Group A. Further analysis of the univariate RR data is limited to inferences at this aggregate data level.

Measurements of individual sensitive behaviors require support by multivariate randomized item-response data. The purpose of this chapter is to give an overview of item response theory models modified such that they are suitable for the analysis of multivariate RR data. The general class of such models is referred to as randomized item-response theory (RIRT) models (Fox 2005; Fox & Wyrick 2008) or item randomized-response (IRR) models (Böckenholt & van der Heijden, 2007). Different RIRT models for binary, ordinal, and mixed responses are presented. Furthermore, it is shown how to extend these models to handle non-compliance, i.e., when respondent do not follow the RR instructions (Clark & Desharnais, 1998), as well as allow for measurement of multidimensional constructs. Our compensatory multidimen-

sional modeling approach generalizes the noncompensatory model by Böckenholt and van der Heijden (2007), who considered multiple item bundles each measuring a specific construct given binary response data.

1.2 Presentation of the Models

In Warner's (1965) approach, a randomizing device (e.g., die, spinner) is required to make the choice between two logically opposite questions. The setup guarantees the confidentiality of each individual response, which cannot be related to either of the opposite questions. Greenberg et al. (1969) proposed a more general unrelated question technique, where the outcome of the randomizing device controls the choice between a sensitive question and an irrelevant unrelated question.

Edgell, Himmelfarb, and Duchan (1982) generalized the procedure by introducing an additional randomizing device to generate the answer on the unrelated question. The responses are then completely protected since it becomes impossible to infer whether they are answers to the sensitive question or forced answers generated by the randomizing device. Let the randomizing device select the sensitive question with probability ϕ_1 and a forced response with probability $1-\phi_1$. The latter is supposed to be a success with probability ϕ_2. Let U_{pi} denote the randomized response of person $p = 1,\ldots,P$ to item $i,1,\ldots,I$. Consider a success a positive response (score one) to a question and a failure a negative response (score zero). Then, the probability of a positive randomized response is represented by

$$P\{U_{pi}=1;\phi_1,\phi_2\} = \phi_1 P\left\{\widetilde{U}_{pi}=1\right\} + (1-\phi_1)\phi_2 \tag{1.1}$$

where $\widetilde{U}_{pi}$ is the underlying response, which is referred to as the true response of person p to item i when directly and honestly answering the question.

For a polytomous randomized response, let $\phi_2(a)$ denote the probability of a forced response in category a for $a = 1,\ldots,A_i$ such that the number of response categories may vary over items. The probability of a randomized response of individual p in category a of item i is given by,

$$P\{U_{pi}=a;\phi_1,\phi_2\} = \phi_1 P\left\{\widetilde{U}_{pi}=a\right\} + (1-\phi_1)\phi_2(a). \tag{1.2}$$

It follows that the forced randomized response model is a two-component mixture model, with the first component modeling the responses to the sensitive question and the second component modeling the forced responses. The mixture probabilities are controlled by the randomizing device. When $\phi_1 > .5$, the randomized response data contain sufficient information to make inferences from the responses.

1.2.1 Randomized IRT Models

In a multivariate setting, multiple items are used to measure an individual latent variable or construct (e.g., alcohol dependence; academic fraud) from multiple correlated randomized item responses. In this setting, the characteristics of the randomizing device are allowed to vary over items. For example, although they relate to the same sensitive latent variable, the sensitivity of items may vary. The variation in sensitivity can then be controlled by adjusting the randomizing device properties. But this option will not be further discussed here.

In randomized IRT modeling, the goal is to model the true item responses, $\widetilde{\mathbf{U}}$, which are latent because they are randomized before being observed. For dichotomous response data, the two-parameter (2PL) normal-ogive model defines the probability of a positive response given the sensitive latent construct θ_p and item discrimination and difficulty parameter a_i and b_i, respectively, which is given by

$$\pi_{pi} = P\left\{\widetilde{U}_{pi} = 1; \theta_p, a_i, b_i\right\} = \Phi\left(a_i\left(\theta_p - b_i\right)\right), \tag{1.3}$$

where $\Phi(.)$ denotes the cumulative normal distribution function.

For polytomous responses, the probability of a response in category a of person p is supposed to be given by

$$\begin{aligned} \pi_{pi}(a) &= P\left\{\widetilde{U}_{pi} = a; \theta_p, a_i, \mathbf{b}_i\right\} \\ &= \Phi\left(a_i\left(\theta_p - b_{i,(a-1)}\right)\right) - \Phi\left(a_i\left(\theta_p - b_{i,a}\right)\right) \end{aligned} \tag{1.4}$$

where vector $\mathbf{b}_i$ contains the threshold parameters of item i , which follow the order restriction: $b_{i1} < \ldots < b_{iA}$ for response alternatives $a = 1, \ldots, A$ (for more on this type of graded response model, see Samejima, vol. 1, chap. 6).

The last model can be extended to deal with questionnaires with items that measure multiple sensitive constructs. Let the multidimensional vector $\boldsymbol{\theta}_i$ of dimension D denote these constructs. Then, the probability of a true response in category a is

$$\begin{aligned} \pi_{pi}(a) &= P\left\{\widetilde{U}_{pi} = a; \boldsymbol{\theta}_p, \boldsymbol{a}_i, \boldsymbol{b}_i\right\} \\ &= \Phi\left(\boldsymbol{a}_i^t\left(\boldsymbol{\theta}_p - b_{i,(a-1)}\right)\right) - \Phi\left(\boldsymbol{a}_i^t\left(\boldsymbol{\theta}_p - b_{i,a}\right)\right) \end{aligned} \tag{1.5}$$

where the vector of discriminations (factor loadings) of dimension D specifies the weights for each underlying dimension.

The models in (1.3)-(1.5) can be embedded in a randomized response modeling framework. For example, for two-parameter normal-ogive model for the true responses, the overall model becomes

$$P\left\{U_{pi} = 1; \theta_p, a_i, b_i\right\} = \phi_1 P\left\{\widetilde{U}_{pi} = 1; \theta_p, a_i, b_i\right\} + (1 - \phi_1)\phi_2. \tag{1.6}$$

Thus, by combining the randomized response technique with an IRT model, latent individual sensitive traits can be measured given observed randomized item responses. A major advantage of using an IRT model is its separation of item parameters and person parameters. Consequently, it can be used to interpret individual differences on the latent trait that is measured, allows for more complex test designs, and handles measurement error at the individual level.

1.2.2 Noncompliant Behavior

Despite the protection of privacy offered by randomized response techniques, some respondents may still show noncompliant behavior and consistently select the least self-incriminating response and completely ignore the randomized response instructions. Clark and Desharnais (1998) proposed a method to estimate the extent of noncompliance using two sampled groups each confronted with a different randomized response designs. Böckenholt and van der Heijden (2007) and Cruyff, van den Hout, van der Heijden, and Böckenholt (2007) proposed the use of a two-component latent class model, where one group consists of respondents that follow the randomized response instructions and a second group of respondents does not follow them.

RIRT modeling can be extended to account for noncompliance. In order to do so, let a binary latent class variable be $G_{pi} = 1$ when person p responds to item i in a noncompliant (self-protective) way and $G_{pi} = 0$ when p responds in a compliant way. Then, the randomized item response model in Equation (1.6) is

$$P\{U_{pi} = 0\} = P\{G_{pi} = 0\}\, P\{U_{pi} = 0; \theta_p, a_i, b_i\} + P\{G_{pi} = 1\}\, I(U_{pi} = 0).$$

where $I(U_{pi} = 0)$ equals one when the answer to item i of respondent p is zero and equals zero otherwise. This mixture model consists of a randomized item response model for the compliant class but a different model for the noncompliant class. Inferences are made from the responses by the compliant class, which requires information about the behavior of the respondents. That is, the assumption of an additional response model for G_{pi} is required (e.g., De Jong et al., 2010; Fox, 2010).

1.2.3 Structural Models for Sensitive Constructs

Respondents are usually independently sampled from a population, and a normal distribution is often used to describe the distribution of the latent variable. If so, the population model for the latent person variable is

$$\theta_p \sim N(\mu_\theta, \sigma^2_\theta)$$

For more complex sampling designs, respondents can be clustered, and the model for the population distribution needs to account for the dependencies between respondents in the same cluster. As described, among others, by Fox (2010) and Fox and Glas (2001; vol. 1, chap. 24), a multilevel population distribution for the latent person parameters needs to be defined. Let θ_{pj} denote the latent parameter of person p in group j ($j = 1, \ldots, J$). The population distribution becomes

$$\begin{aligned} \theta_{pj} &\sim N\left(\beta_j, \sigma^2_\theta\right) \\ \beta_j &\sim N\left(\mu_\theta, \tau^2_{00}\right). \end{aligned}$$

Or, for the multidimensional case,

$$\boldsymbol{\theta}_p \sim N\left(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta\right),$$

where the covariance matrix of dimension D specifies the within-person correlations.

This multidimensional model can also be extended to include a multilevel setting, but this case will not be discussed. Also, to explain variation between persons in latent sensitive measurements, explanatory variables at the level of persons and/or groups can also be included. Finally, variation in item parameters can also be modeled as described in De Boeck and Wilson (2004; vol. 1, chap. 33) and De Jong et al. (2010).

1.3 Parameter Estimation

A fully Bayesian estimation method with MCMC sampling from the posterior distribution of the parameters is presented. The method requires prior distributions for all model parameter. Non-informative inverse gamma priors are specified for the variance components. An inverse Wishart prior is specified for the covariance matrix. Normal and lognormal priors are specified for the difficulty and discrimination parameters, respectively. A uniform prior is specified for the threshold parameters while accounting for the order constraint.

Following the MCMC sampling procedure for item randomized-response data in Fox (2005, 2010), Fox and Wyrick (2008), and De Jong et al. (2010), a fully Gibbs sampling procedure is developed which consists of a complex data augmentation scheme: (i) sampling of latent true responses, $\widetilde{\boldsymbol{U}}$; (ii) sampling latent continuous response data, $\boldsymbol{Z}$; and (iii) sampling latent class membership $\boldsymbol{G}$. The item response model parameters and structural model parameters are sampled in a straightforward way given the continuous augmented data, as described by Fox (2010) and Johnson and Albert (2001).

Omitting conditioning on $G_{pi} = 0$ for notational convenience, the procedure is described for latent response data generated only for responses belonging to the compliant class. A probabilistic relationship needs to be defined between the observed randomized response data and the true response data. To do so, define $H_{pi} = 1$ when the randomizing device determines that person i answers item i truthfully and $H_{pi} = 0$ when a forced response is generated. It follows that the conditional distribution of a true response a given a randomized response a' is given by

$$\begin{aligned} P\left\{\widetilde{U}_{pi} = a' \mid U_{pi} = a\right\} &= \frac{P\left\{\widetilde{U}_{pi} = a', U_{pi} = a\right\}}{P\{U_{pi} = a\}} \\ &= \frac{\sum_{l\in\{0,1\}} P\left\{\widetilde{U}_{pi} = a', U_{pi} = a \mid H_{pi} = l\right\} P\{H_{pi} = l\}}{\sum_{l\in\{0,1\}} P\{U_{pi} = a \mid H_{pi} = l\} P\{H_{pi} = l\}}, \end{aligned}$$

where $a, a' = \{0,1\}$ and $\{1,2,\ldots,A_i\}$ for binary and polytomous responses, respectively.

For binary responses, π_{pi} in Equation 1.3 defines the probability of a success.

Subsequently, the latent responses are Bernoulli distributed,

$$\begin{aligned}\widetilde{U}_{pi} \mid U_{pi}=1,\pi_{pi} &\sim B\left(\lambda=\frac{\pi_{pi}\,(p_1+p_2(1-p_1))}{p_1\pi_{pi}+p_2(1-p_1)}\right),\\ \widetilde{U}_{pi} \mid U_{pi}=0,\pi_{pi} &\sim B\left(\lambda=\frac{\pi_{pi}(1-p_1)(1-p_2)}{1-(p_1\pi_{pi}+p_2(1-p_1))}\right).\end{aligned}$$

For polytomous response data, π_{pi} is defined in Equation (1.4) or (1.5), and $\widetilde{U}_{pi}$ given $U_{pi}=a$ is multinomially distributed with cell probabilities

$$\Delta(a)=\frac{\pi_{pi}(a')p_1 I(a=a')+\pi_{pi}(a')(1-p_1)p_2(a)}{\pi_{pi}(a)p_1+(1-p_1)p_2(a)}.$$

Following the data augmentation procedure of Johnson and Albert (2001) and Fox (2010), latent true response data are sampled given the augmented dichotomous or polytomous true response data.

The latent class memberships, G_{pi}, are generated from a Bernoulli distribution. Let $Y_{pi}=0$ define the least self-incriminating response, then the success probability of the Bernoulli distribution can be expressed as

$$\frac{P\{G_{pi}=1\}I(Y_{pi}=0)}{P\{G_{pi}=0\}P\{Y_{pi}=0\mid\theta_p,a_i,b_i\}+P\{G_{pi}=1\}I(Y_{pi}=0)}$$

where a Bernoulli prior is usually specified for the class membership variable G_{pi}.

Given the augmented data, class memberships, true responses, and latent true responses, all other model parameters can be sampled using a full Gibbs sampling algorithm. The full conditionals can be found in the MCMC literature for IRT (e.g., Junker, Patz, & Vanhoudnos, vol. 2, chap. 15).

1.4 Model Fit

A Bayesian residual analysis can be performed to evaluate the fit of the model. Residual analysis for binary and polytomous item response models has been suggested by De Jong et al. (2010), Fox (2010), Geerlings, Glas, and van der Linden (2011), and Johnson and Albert (2001). Posterior distributions of the residuals can be used to evaluate their magnitude and make probability statements about them. Bayesian residuals are easily computed as by-products of the MCMC algorithm, and they can be summarized to provide information about specific model violations. For instance, sums of squared residuals can be used as a discrepancy measure for evaluating person or item fit. The extremeness of the observed discrepancy measure can be evaluated using replicated date generated under their posterior predictive distribution. Likewise, the assumption of local independence and unidimensionality can be checked using appropriate discrepancy measures. For an introduction to posterior predictive

checks, see Sinharay (vol. 2, , chap. 19). Studies of different posterior predictive checks for Bayesian IRT models are reported in Glas and Meijer (2003), Levy, Mislevy, and Sinharay (2009), Sinharay, Johnson, and Stern (2006), and Sinharay (2006).

1.5 Empirical Example

In a study of alcohol-related expectancies and problem drinking, responses to thirteen items of the College Alcohol Problem Scale (CAPS; O'Hare, 1997) and four items of the Alcohol Expectancy questionnaire (AEQ; Brown, Christiansen, & Goldman, 1987) were analyzed. The goal was to measure the sensitive constructs underlying both scales using multidimensional item response theory. Furthermore, it was investigated whether the randomized response technique improved the accuracy of the self-reports obtained by direct questions.

1.5.1 CAPS and AE Questionnaire

As an initial screening instrument, the CAPS instrument was developed to measure drinking problems among youth. Its items covered socio-emotional problems, such as hangovers, memory loss, nervousness, and depression, as well as community problems, such as drove under the influence, engaged in activities related to illegal drugs, problems with the law. The questionnaire items are given in Appendix A. Self-reported information about negative consequences of drinking is likely to be biased due to socially desirable responding. Consequently, the survey was expected to lead to refusals to respond and responses given to conceal undesirable behavior. Therefore, a randomized response technique was used to improve both the cooperation by the respondents and the accuracy of their self reports.

The AEQ measures the degree of expectancies associated with drinking alcohol. Alcohol-related expectancies are known to influence alcohol use and behavior while drinking. The adult form of the AEQ consisted of 90 items and covers six dimensions. But in the study the focus was on alcohol-related sexual enhancement expectancies. The items covering sexual enhancement expectancies are given in Appendix A. The data were collected on a five-point ordinal scale, ranging from one (almost never) to five (almost always).

The CAPS data were re-analyzed by Fox and Wyrick (2008), who used a unidimensional randomized item response model to measure general alcohol dependence. Although the model described the data well, CAPS was developed by O'Hare (1997) to measure different psychosocial dimensions of problem drinking among college students. Two of the dimensions, socio-emotional and community problems, were identified by analysis. Together, they explained more than 60% of the total variance of the responses. In the present study, a multidimensional modeling approach was carried out to investigate whether the CAPS data supported the measurement of multiple sensitive constructs given randomized responses. Besides, the multidimensional

model was also used to jointly analyze the CAPS and AEQ data for the relationships between the multiple factors they measure. Finally, the effects of the randomized response technique on the measurement of their factors was analyzed jointly.

1.5.2 Data

A total of seven hundred ninety-three students from four local colleges/universities, Elon University (N=495), Guilford Technical Community College (N=66), University of North Carolina (N=166), and Wake Forest University (N=66), participated in the survey study in 2002. Both the CAPS and AEQ items were administered to them and their age, gender, and ethnicity was recorded. It was logistically not possible to randomly assign students to the direct questioning (DQ) or the randomized response (RR) condition. However, it was possible to randomly assign classes of five to ten participants to one of the conditions.

A total of 351 students was assigned to the DQ condition. They served as the control group and were instructed to answer the questionnaire as they normally would. A total of 442 students in the RR condition received a spinner to assist them in completing the questionnaire. For each item, the spinner was used as a randomizing device which determined whether to answer honestly or to give a forced response. According to a forced response design, the properties of the spinner were set such that an honest answer was requested with a probability of .60 and a forced response with a probability of .40. When a forced response was to be given, each of the five possible responses had a probability of .20.

1.5.3 Model Specification

The following multidimensional randomized item response model was used to analyze the data,

$$\begin{aligned} P(Y_{pi} = a \mid \boldsymbol{\theta}_p, \boldsymbol{a}_i, \boldsymbol{b}_i) &= p_1\pi_{pi} + (1-p_1)p_2(a) \\ \pi_{pi} &= \Phi\left(\mathbf{a}_i^t\left(\boldsymbol{\theta}_p - b_{i,(a-1)}\right)\right) - \Phi\left(\mathbf{a}_i^t\left(\boldsymbol{\theta}_p - b_{i,a}\right)\right), \quad (1.7) \\ \boldsymbol{\theta}_p &\sim N\left(\boldsymbol{\mu}_{\theta,p}, \boldsymbol{\Sigma}_\theta\right) \\ \boldsymbol{\mu}_{\theta,p} &= \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 RR_p \end{aligned}$$

for $a = 1,\ldots,5$ and $i = 1,\ldots,17$. As just indicated, in the forced randomized response sampling design, $p_1 = .60$ and $p_2(a) = .20$, for $a = 1,\ldots,5$, whereas for the direct-questioning conditioning $p_1 = 1$. The explanatory variable RR_p was equal to one when student p belonged to the RR group and equal to zero otherwise. The factor loadings, **a,** and item thresholds were assumed to be independent of the questioning technique.

Following Béguin and Glas (2001), the model was identified by fixing the mean score for each dimension, such that $\boldsymbol{\beta}_0 = \mathbf{0}$, while the variance components for each factor to set equal to one. To avoid the so-called rotational variance, one item was assigned uniquely to each of the Q dimensions.

The MCMC algorithm was used to estimate simultaneously all model parameters using $50,000$ iterations, with a burn-in period of $10,000$ iterations.

1.5.4 Results

In Table 1.1, the estimated factor loadings for a three-factor of the multidimensional RIRT model in (1.7) are given. The factor loadings were standardized by dividing each of them by the average item loading. Furthermore, for each factor the sign of the loadings was set such that a higher latent score corresponded to a higher observed score. To avoid label switching, Items 1, 5, and 14 were allowed to have one free non-zero loading, so that each of these items represented one factor.

Items 1-4, 6,8, and 9 were positively associated with the first factor and had factor loadings higher than .60. This first factor represents drinking-related socio-emotional problems, including depression, anxiety, and troubles with family. These problems increased with alcohol consumption. Some of the items also loaded on the two other factors.

The second factor (community problems) covered Items 5,7,and 10-13, with loadings higher than .60, except for Item 12. In the literature, Item 12 has been associated with factor community problems, but in our analysis the item also related to the other factors, most strongly to the third. This second factor covers acute physiological effects of drunkness together with illegal and potentially dangerous activities (e.g., driving under the influence).

As expected, Items 14-17 were associated with a third factor, which represented alcohol-related sexual enhancement expectancies. These expectancies increased with alcohol consumption but, given their negative loadings on the other two factors, slightly reduced the socio-emotional and community problems.

The multivariate latent factor model was extended with an explanatory variable denoted as RR, which indicated when a student was assigned to the RR (RR=1) or the DQ condition (RR=0). In addition, an indicator variable was included, which was set equal to one when the respondent was a female. Both explanatory variables were used for each factor. The RIRT model was further extended with a multivariate population model for all factors.

In Table 1.2, the parameter estimates of the three-factor and a two-factor model are given. For the latter, the loadings of Items 1 and 14 were fixed to identify two factors, with one factor representing a composite measure of alcohol-related problems (i.e., socio-emotional and community problems) and the other alcohol-related sexual enhancement expectancies. A moderate positive correlation of .65 between the two factors was found.

The students in the RR condition scored significantly higher on both factors. For the RR group, the average latent scores were .20 and .22 on the composite problem and the alcohol-related expectancy factors, respectively, but both were equal to zero for the DQ group. The RR effect was slightly smaller than that of .23 reported by Fox and Wyrick (2008), who performed a unidimensional RIRT analysis using the CAPS items only. A comparable effect was found for the AEQ scale. Females and males showed comparable scores on both factors.

TABLE 1.1
CAPS-EAQ Scale: Weighted factor loadings for the three-component analysis.

Subscale Items	Three-Factor RIRT Model		
Socio-Emotional problems	Factor 1	Factor 2	Factor 3
1 Feeling sad, blue or depressed	1.00	.00	.00
2 Nervousness or irritability	1.00	.01	-.03
3 Hurt another person emotionally	.96	.27	.10
4 Family problems related to drinking	.82	.56	.14
6 Badly affected friendship	.85	.46	.27
8 Other criticize your behavior	.77	.50	.41
9 Nausea or vomiting	.70	.39	.60
Community problems	Factor 1	Factor 2	Factor 3
5 Spent too much money on drugs	.00	1.00	.00
7 Hurt another person physically	.48	.84	.26
10 Drove under the influence	.43	.74	.53
11 Spent too much money	.59	.66	.47
12 Feeling tired or hung over	.57	.41	.71
13 Illegal activities	.05	.96	.29
Sexual enhancement	Factor 1	Factor 2	Factor 3
14 I often feel sexier	.00	.00	1.00
15 I'm a better lover	-.09	-.12	.99
16 I enjoy having sex more	-.14	-.06	.99
17 I am more sexually responsive	-.17	-.03	.99

In the three-factor model, with the estimated loadings given in Table 1.1, the problems associated with drinking were represented by two factors (i.e., socio-emotional and community problems) and sexual enhancement expectancies by another factor. The randomized response effects were significantly different from zero for all three factors, while the effect on the factor representing community problems related to alcohol use was approximately .32. This was slightly higher than the effects of the other components, which were around .21. It seemed as if the students were less willing to admit to alcohol-related community problems and gave more socially desirable responses than for the other factors.

The male students scored significantly higher than the female students on the factor representing community problems related to alcohol use. That is, male students were more likely to experience alcohol-related community problems than females. This gender effect was not found for the other factors. The estimated effects indicated that the RR-group scored significantly higher in comparison to the DQ-group on each subscale. Although validation data are not available, the RR technique was expected to have led to an improved willingness of the students to answer truthfully, given their random assignment to the direct questioning and randomized response conditions.

Finally, the three factors yielded moderate positive correlations, as shown in Table 1.2. The factors community and socio-emotional problems correlated positively with sexual enhancement expectancies due to alcohol use. In line with the alcohol expectancy theory, more positive expectancies of alcohol use lead to more positive

drinking experiences, which in turn lead to more positive expectancies. Here, an increased expectancy of sexual enhancement stimulates alcohol use, which leads to more socio-emotional and community problems.

1.6 Discussion

Response bias is a serious threat to any research that uses self-report measures. Subjects are often not willing to cooperate or to provide honest answers to personal, sensitive questions. The general idea is that by offering confidentiality, respondents will become more willing to respond truthfully. Warner's (1965) randomized response technique was developed to ensure such confidentiality.

Our multivariate extension of the technique still masks the responses to the items but enables us to estimate item characteristics and measure individual differences in sensitive behavior. The models can handle both dichotomous and polytomous responses to measure both unidimensional or multidimensional sensitive constructs. In the empirical example above, a forced randomized response design was used to collect the data, but other options are available. Our RIRT models are easily adapted to a specific choice of response design.

In order to improve the cooperation of the respondents, both from an ethical and professional point of view, they should be informed about the levels of information that can and cannot be inferred from randomized item responses. The outcome of the randomization device is only known to the respondent, which protects them at the level of the individual items.

The randomized response technique also has some disadvantages. The use of a randomization device makes the procedure more costly, and respondents have to trust the device. Respondents also have to understand the procedure to recognize and appreciate the anonymity they guarantee. Recently, Jann, Jerke, and Krumpal (2012), Tan, Tian, and Tang (2009), and Coutts and Jann (2011) proposed nonrandomized response techniques to overcome the inadequacies of the randomized response technique and tested their proposals empirically. The main idea of their so-called triangular and crosswise technique is to ask respondents a sensitive and a nonsensitive question and let them indicate whether the answers to the questions are the same (both 'Yes' or both 'No') or different (one 'Yes' and the other 'No'). Such a joint answer to both questions does not reveal the respondent's true status. The distribution of answers to the nonsensitive question has to be known and supports the measurement of the population prevalence on the sensitive question. These nonrandomized methods are designed to make inferences at an aggregate data level. Extensions are required to collect multivariate sensitive items responses that will support the measurement of sensitive constructs. In fact, more research is needed to explore the full potential of nonrandomized response techniques for the analysis of individual sensitive constructs.

TABLE 1.2
CAPS-EAQ scale: Parameter estimates of two- and three-component randomized item-response model.

	Two Factor		Three Factor	
Parameter	Mean	SD	Mean	SD
Fixed Effects				
Socio-Emotional/Community				
γ_{11} (RR)	.20	.09	.21	.10
γ_{21} (Female)	.01	.06	.05	.07
Sexual enhancement expectancy				
γ_{12} (RR)	.22	.06	.21	.07
γ_{22} (Female)	.03	.04	.06	.05
Community				
γ_{13} (RR)			.32	.10
γ_{23} (Female)			$-.30$	.09
Variance Parameters	Mean	SD	Mean	SD
$\Sigma_{\theta_{11}}$	.96	.05	.98	.05
$\Sigma_{\theta_{12}}$	.65	.07	.55	.06
$\Sigma_{\theta_{13}}$			.38	.08
$\Sigma_{\theta_{22}}$	.98	.05	1.06	.05
$\Sigma_{\theta_{23}}$			.42	.08
$\Sigma_{\theta_{33}}$			.99	.07
Information Criteria				
-2log-likelihood		20622		19625

Acknowledgement

The author thanks Cheryl Haworth Wyrick for providing the data from the study on alcohol use and abuse by college students.

References

Anglin, D., Hser, Y., & Chou, C. (1993). Reliability and validity of retrospective behavioral self-report by narcotics addicts. *Evaluation Review, 17*, 91-108.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation of multidimensional IRT models. *Psychometrika, 66*, 541-562.

Böckenholt, U. & van der Heijden, P. G. M. (2007). Item randomized–response models for measuring noncompliance: Risk–return perceptions, social influences, and self-protective responses. *Psychometrika, 72*, 245–62.

Brown, S. A., Christiansen, B. A., & Goldman, A. (1987). The alcohol expectancy questionnaire: An instrument for the assessment of adolescent and adult alcohol expectancies. @Author: add journal title@*, 48*, 483-491.

Clark, S. J. & Desharnais, R. A. (1998). Honest answers to embarrassing questions: Detecting cheating in the randomized response model. *Psychological Methods, 3*, 160-168.

Coutts, E. & 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.

Cruyff, M. J. L. F., van den Hout, A., van der Heijden, P. G. M., & Böckenholt, U. (2007). Log-linear randomized-response models taking self-protective response behavior into account. *Sociological Methods and Research, 36*, 266-282.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

De Jong, M. G., Pieters, R. & Fox, J.-P. (2010). Reducing social desirability bias through item randomized response: An application to measure underreported desires. *Journal of Marketing Research, 47*, 14-27.

Edgell, S. E., Himmelfarb, S., & Duchan, K. L. (1982). Validity of forced Responses in a randomized response model. *Sociological Methods and Research, 11*, 89-100.

Fox, J. A., & Tracy, P. E. (1986). *Randomized response: A method for sensitive surveys*. London: Sage.

Fox, J.-P. (2005). Randomized item response theory models. *Journal of Educational and Behavioral Statistics, 30*, 189-212.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York: Springer.

Fox, J.-P. & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*, 269-286.

Fox, J.-P., & Wyrick, C. (2008). A mixed effects randomized item response model. *Journal of Educational and Behavioral Statistics, 33*, 389-415.

Geerlings, H., Glas, C. A. W., & van der Linden, W. J. (2011). Modeling rule-based item generation. *Psychometrika, 76*, 337-359.

Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement, 27*, 217-233.

Greenberg, B. G., Abul-Ela, A.-L. A., Simmons, W. R., & Horwitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association, 64*, 520-539.

Jann, B., Jerke, J., & Krumpal, I. (2012). Asking sensitive questions using the crosswise model: An experimental survey measuring plagiarism. *Public Opinion Quarterly, 76*, 1-18.

Johnson, V. E. & Albert, J. H. (2001). *Ordinal data modeling*. New York: Springer.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement, 33*, 519-537.

O'Hare, T. (1997). Measuring problem drinking in first time offenders: Development and validation of the college alcohol problem scale (CAPS). *Journal of Substance Abuse Treatment, 14*, 383-387.

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models.*British Journal of Mathematical and Statistical Psychology,59*, 429-449.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30*, 298-321.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology.* San Francisco, CA: Jossey-Bass.

Tan, M. T., Tian, G.-L., & Tang, M.-L. (2009). Sample surveys with sensitive questions: A nonrandomized response approach. *The American Statistician, 63*, 9-16.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. New York: Cambridge University Press.

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: The impact of data collection, question format, and question technique. *Public Opinion Quarterly, 60*, 275-304.

van der Heijden, P. G. M., van Gils, G., Bouts, J., & Hox, J. J. (2000). 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, 28*, 505-537.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 60*, 63-69.

Appendix A: CAPS-AEQ Questionnaire

CAPS: Socio-emotional and community problems
How often (almost always (5), often (4), sometimes (3), seldom (2), almost never (1)) have you had any of the following problems over the past years as a result of drinking too much alcohol?

1. Feeling sad, blue or depressed.
2. Nervousness or irritability
3. Hurt another person emotionally
4. Family problems related to your drinking
5. Spent too much money on drugs
6. Badly affected friendship or relationship
7. Hurt another person physically
8. Caused other to criticize your behavior
9. Nausea or vomiting
10. Drove under the influence
11. Spent too much money on alcohol
12. Feeling tired or hung over
13. Illegal activities associated with drug use

AEQ: Sexual enhancement

14 I often feel sexier after I've had a couple of drinks
15 I'm a better lover after a few drinks
16 I enjoy having sex more if I've had some alcohol
17 After a few drinks, I am more sexually responsive