

Randomized Item Response Theory Models

J.-P. Fox

University of Twente

The randomized response (RR) technique is often used to obtain answers on sensitive questions. A new method is developed to measure latent variables using the RR technique because direct questioning leads to biased results. Within the RR technique is the probability of the true response modeled by an item response theory (IRT) model. The RR technique links the observed item response with the true item response. Attitudes can be measured without knowing the true individual answers. This approach makes also a hierarchical analysis possible, with explanatory variables, given observed RR data. All model parameters can be estimated simultaneously using Markov chain Monte Carlo. The randomized item response technique was applied in a study on cheating behavior of students at a Dutch University. In this study, it is of interest if students' cheating behavior differs across studies and if there are indicators that can explain differences in cheating behavior.

Keywords: analysis of variance, item response theory model, Markov chain Monte Carlo (MCMC), random effects, randomized response

The collection of data through surveys on highly personal and sensitive issues may lead to answering refusals and false responses, making inferences difficult. Obtaining valid and reliable information depends on the cooperation of the respondents, and the willingness of the respondents depends on the confidentiality of their responses. Warner (1965) developed a data collection procedure, the randomized response (RR) technique, that allows researchers to obtain sensitive information while guaranteeing privacy to respondents. For example, a randomizing device is used to select a question from a group of questions and the respondent answers the selected question. The respondent is protected because the interviewer will not know which question is being answered. Warner's and related approaches were specifically developed to hide answers of the respondents and to estimate proportions and related confidence intervals in the population.

In some applications, randomized response data can be hierarchically structured, and there is an interest in group differences regarding some sensitive individual characteristic. One could think of an application where it is of interest to know if cheating behavior differs across faculties or if social security fraud is more likely to appear in groups with certain characteristics. However, the usual randomized response models do not allow a hierarchical data analysis of the RR data. Statistical methods for hierarchically structured data, for example analysis of variance (ANOVA) or multilevel analysis, cannot be applied because the true individual

responses are unknown. The true item responses are defined as the item responses that would have been obtained when directly asking sensitive questions given truthful answers of the respondents. In most situations, it can be assumed that direct questioning, for obtaining sensitive information, leads to biased results.

It is shown that this problem can be circumvented when item response theory (IRT) is used for modeling the true responses to a set of items. An IRT model specifies the probability of a respondent achieving a score on an item, as a function of the respondent's continuous valued unobservable (latent) attitude and item parameters (Lord & Novick, 1968). The RR model relates the observed RR data with the true item response data. The proposed method enables researchers to perform a hierarchical analysis of some underlying characteristic or attitude measured by a set of randomized item responses.

It is well known that data collection by means of direct questioning often results in biased estimates when asking sensitive questions. Furthermore, the RR technique estimators of the population proportions have larger standard errors than estimators from direct questioning. Therefore, Scheers and Dayton (1988) developed a covariate randomized response model. The RR technique estimation is improved by using a covariate that correlates with the individual latent characteristic or attitude. This technique is easily incorporated in the proposed method using IRT. That is, covariates concerning individual or group characteristics can be taken into account in the estimation of the true individual sensitive attitudes given the randomized responses. However, it is also possible the other way around, to explore characteristics that can be held responsible for individual differences in attitudes. Therefore, the differential influence of individual and group characteristics on sensitive characteristics can be investigated.

This article can be divided into three parts. In the first part, the proposed randomized item response theory (RIRT) model is described for the Warner technique and the unrelated-question technique (Greenberg, Abul-Ela, Simons, & Horvitz, 1969). A description is given of the incorporation of covariates and the hierarchical RIRT model. The second part shows an application of the hierarchical RIRT model to data on cheating behavior of students at a University in the Netherlands. In the final part, a Markov chain Monte Carlo (MCMC) estimation procedure is described for estimating simultaneously all parameters of the hierarchical RIRT model. As a result, the extra variance induced by the RR technique is taken into account in the estimation of the other parameters.

The Randomized Item Response Model

The commonly used RR models, Warner's model and the unrelated-question model, are extended with an IRT model to give rise to a hierarchical analysis of the underlying attitude.

The Warner Model

In the data collection procedure, a random device is used in such a way that a respondent answers one of two randomly selected questions. Furthermore, the

respondent does not reveal the question that has been selected. When answering a set of items, each item is presented in a positively and negatively worded form. Let π denote the true probability of giving a positive response ($Y = 1$) and p denote the known probability of selecting a particular question of an item based on the randomization device. It follows that

$$P(Y = 1) = p\pi + (1 - p)(1 - \pi). \quad (1)$$

From a sample of responses, the true probability of a positive response in the population, π , can be estimated (Warner, 1965) given the observed responses. In a different perspective, it could be stated that the true item response, denoted as \tilde{y} , that would have been obtained by means of direct questioning is missing. Only the observed response y is obtained because of the interference of the randomization device.

IRT models are commonly used for measuring attitudes or abilities given item responses. Obviously, an IRT model can be used to measure a (sensitive) characteristic if the true item responses are observed. Suppose that respondents, indexed ij ($i = 1, \dots, n_j, j = 1, \dots, J$), are nested in groups, indexed ($j = 1, \dots, J$). The true item responses to the $k = 1, \dots, K$ items are captured in a matrix \tilde{y} . It is assumed that the set of items are composed to measure some underlying attitude θ . According to a two-parameter (normal ogive) IRT model: the probability of a true positive response on item k of a respondent, indexed ij , with attitude level θ_{ij} , is given by

$$P(\tilde{Y}_{ijk} = 1 \mid \theta_{ij}, a_k, b_k) = \Phi(a_k\theta_{ij} - b_k), \quad (2)$$

where a_k and b_k are the item parameters measuring the discriminating power and difficulty of item k , and Φ denotes the standard cumulative distribution function. An item with high discriminating power results in high scores of respondents with high attitude level and low scores of respondents with low attitude level. The respondents score about the same when the discriminating power is low. Respondents score poorly on items with a high difficulty parameter and well on items with a low difficulty parameter.

The normal ogive model, Equation 2, is easily integrated in Warner's model. It follows that the probability of a positive response on item k for a student indexed ij is

$$P(Y_{ijk} = 1) = p\pi_{ijk} + (1 - p)(1 - \pi_{ijk}), \quad (3)$$

where $\pi_{ijk} = P(\tilde{Y}_{ijk} = 1 \mid \theta_{ij}, a_k, b_k)$. In this randomized item response theory (RIRT) model, Equation 3, the true nonobserved item responses are modeled by an IRT model, and they are linked to the observed item responses via Warner's RR technique.

The whole concept becomes clearer through a probability tree. In Figure 1 it is shown how a true response to item k of a person indexed ij leads to an observed item response according to the RIRT model based on Warner's RR technique. First, the value of the unobserved true response \tilde{y}_{ijk} depends on the value of the latent attitude, θ_{ij} , and the item parameters. Second, the observed item response

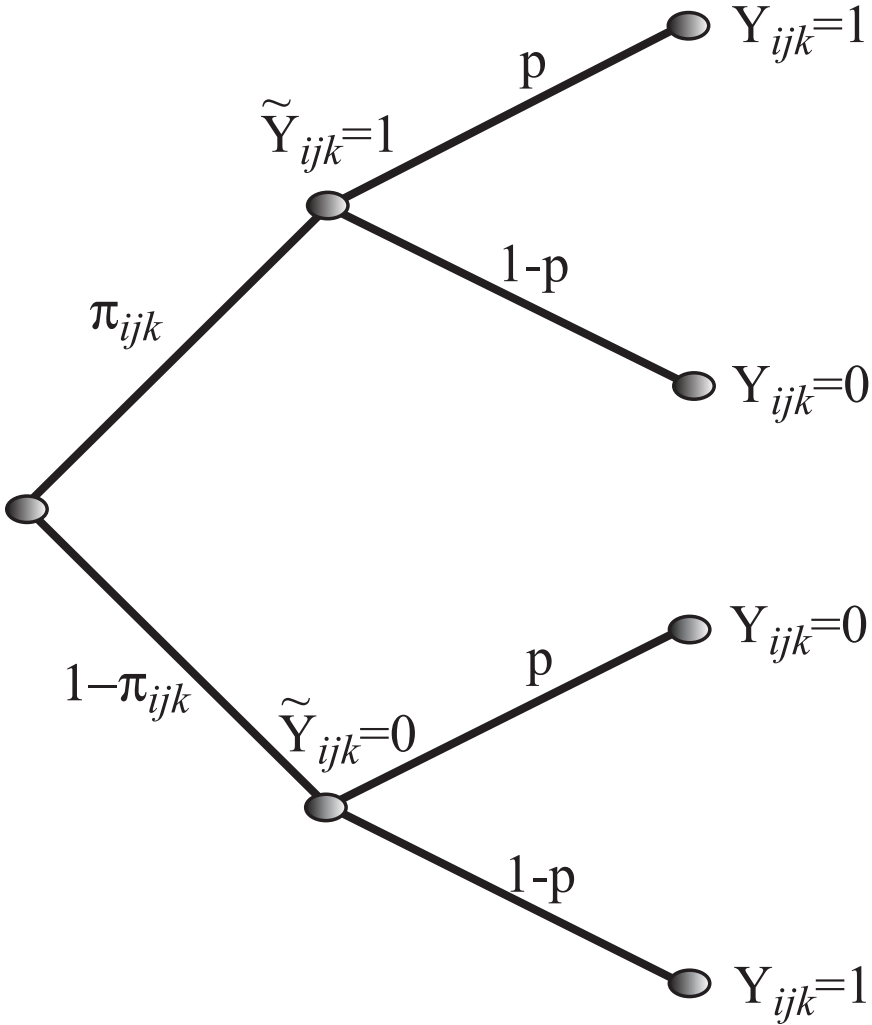


FIGURE 1. Probability tree of the RIRT model using Warner's RR technique.

depends on a probability p attributable to the interference of the randomization device. For example, the respondent, with attitude level θ_{ij} , gives a positive true (latent) response on item k with probability $\Phi(a_k\theta_{ij} - b_k)$. A negative response is observed with a probability $1 - p$.

The Unrelated-Question Model

In the unrelated-question model, the respondent replies to one of the two questions, but one of the questions is completely innocuous and unrelated to the under-

lying characteristic. Greenberg et al. (1969) assumed that the respondent might be more truthful when using the unrelated-question technique. In the most simple form, the outcome of the unrelated question is known, and the corresponding proportion of positive answers is already known. So, the object is to estimate the true proportion of positive responses on the sensitive question. This follows in a natural way in the forced alternative method (Fox & Tracy, 1986). The randomized device determines whether the respondent is forced to answer positively, negatively, or to answer the sensitive question. For example, in the study described below, concerning cheating behavior of students at a Dutch University two dice were used. The respondents were asked to roll two dice and answer “yes” if the sum of the outcomes were 2 or 3, answer the sensitive question if the sum were between 3 and 11, and answer “no” if the sum were 12. Let p_1 be the probability that the respondent has to answer the sensitive question and p_2 be the probability of a forced positive response given that a forced response has to be given. Here, $p_1 = 3/4$ and $p_2 = 2/3$. So, the probability of observing a positive response on item k from respondent ij , that is $Y_{ijk} = 1$, is

$$P(Y_{ijk} = 1) = 3/4\pi_{ijk} + 1/4 \cdot 2/3. \quad (4)$$

The probability π_{ijk} represents the probability on a positive response when directly asking the sensitive question. Again, an IRT model can be used to model the true item responses.

In Figure 2, a probability tree is given that presents the relationship between the true response and the observed response. For example, there are three ways of observing a positive response. With a probability π_{ijk} , a positive response is given on the sensitive question, with probability p_1 , this response is also observed, and with probability $(1 - p_1)p_2$, a forced positive response is observed. With probability $1 - \pi_{ijk}$ a negative response on the sensitive question is given, but a forced positive response is observed with probability $(1 - p_1)p_2$. The graph displays all possible outcomes with probabilities.

The forced alternative method is comparatively easy for respondents to comprehend. Furthermore, it can be safely assumed that the values of the parameters p_1 and p_2 are known. In the unrelated-question model, it is not always clear if these parameters are known without error or have to be estimated. When asking an unrelated question concerning a characteristic of the respondents, the population under survey may have changed or may be slightly different from a larger population for whom the proportion of positive responses is known.

A Hierarchical RIRT Model

There are at least two interesting extensions of the described RIRT model. In the first, the population proportions of the true positive responses may vary from group to group. This way it can be investigated if the proportions of true positive responses, regarding a particular item, differs per group. In the second extension, the population attitude means may differ. Then, interest is focused on attitude differences between groups.

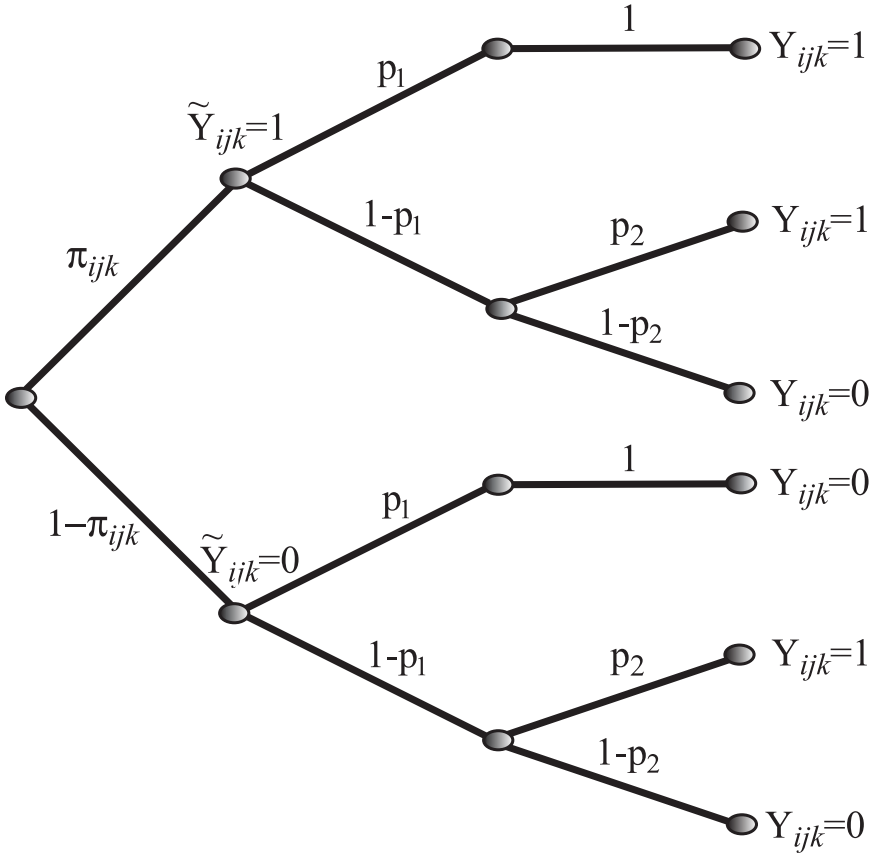


FIGURE 2. Probability tree of the RIRT model using the unrelated-question model.

ANOVA of Population Proportions

It is assumed that the groups are randomly selected from a larger population. Then, the true mean population proportions can be broken down in a group contribution, a random group effect plus a general population mean, and a deviation for each respondent from their group’s contribution. Consider the responses to item k . Let π_{ijk} denote the probability on a positive true item response defined in Equation 2. It follows that

$$\Phi^{-1}(\pi_{ijk}) = \mu_k + \zeta_{jk} + \epsilon_{ijk}, \tag{5}$$

where μ_k is the general mean, considering all responses to item k , and ζ_{jk} is the random group effect. The random effects may be transformed, $\zeta_{jk} = \zeta_{jk} - \bar{\zeta}_{jk}$, where $\bar{\zeta}_{jk}$ is the general mean. It is assumed that $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma_{\epsilon}^2)$ and $\zeta_{jk} \sim \mathcal{N}(0, \sigma_{\zeta}^2)$. That is, the responses of subjects in different groups are independent. The model in

Equation 5 is a random effects model with a probit link for the true Bernoulli response. Notice that the true Bernoulli responses, modeled by an IRT model in Equation 2, are based on the randomized responses. The general mean is the unweighted average of all group means and implies that the random group effect ζ_{jk} has expected value zero. It is possible to extend the model in Equation 5 by introducing individual respondent or group characteristics. They can explain differences between the individual probabilities, or increase the accuracy of the corresponding estimates.

ANOVA of Individual Attitudes

Test or questionnaires are used to measure individual attitudes that are sensitive in nature. Therefore, an RR technique is used because it is expected that certain questions elicit either wrong answers or noncooperation from the respondent. The extension of an RR model with an IRT model includes latent attitudes of the respondents. The second model extension concerns the latent attitudes.

When the respondents are nested in groups, it can be assumed that different attitudes are more alike when they belong to the same group. Assume that the groups are sampled from a larger population of groups. Then, the model equation is

$$\theta_{ij} = \gamma_0 + u_j + e_{ij}, \tag{6}$$

where $u_j \sim \mathcal{N}(0, \sigma_u^2)$, and $e_{ij} \sim \mathcal{N}(0, \sigma_e^2)$. Here, γ_0 is the general mean, and u_j is the random group effect on the attitude of the respondent indexed ij . There are various interesting possibilities. For example, in the real data example described below, it was investigated if group mean attitudes regarding cheating differed across studies. In a fixed effect ANOVA, it can be tested if, for example, males are more willing to commit social security fraud or if male students are more likely to cheat on examination tests given RR data. When individual or group characteristics are available, the RIRT model can be extended to a multilevel model with a latent dependent variable. This model resembles the multilevel IRT model developed by Fox (2004) and Fox and Glas (2001), except that here the observations are obtained via an RR technique. A multilevel model describes the relationships between the “outcome” variable (attitudes, abilities), group characteristics (group size, financial resources), and respondents’ characteristics (achievements, social background). Then, individual and group characteristics can be used to explore differences within and between groups regarding the measured attitudes.

Parameter Estimation

The hierarchical RIRT model contains three components, a randomized response model, $p(\mathbf{Y} | \tilde{\mathbf{Y}})$, that relates the observed item responses with the true underlying item responses assuming that the probabilities concerning the randomization device are known. An item response model, $p(\tilde{\mathbf{Y}} | \boldsymbol{\theta}, \mathbf{a}, \mathbf{b})$ for measuring the underlying attitudes. A structural hierarchical model, $p(\boldsymbol{\theta} | \mathbf{X}, \mathbf{W}, \boldsymbol{\beta})$, where \mathbf{X} and \mathbf{W} are individual and group characteristics, respectively, and $\boldsymbol{\beta}$ are the fixed and/or random

effects. This last component is the target of inference and comprehends the ANOVA or multilevel analysis of the respondents' attitudes.

MCMC estimation (see, e.g., Gelfand & Smith, 1990; Geman & Geman, 1984; Gilks, Richardson, & Spiegelhalter, 1996; Tanner & Wong, 1987) is a powerful tool for estimation in complex models. An MCMC procedure can be applied to estimate simultaneously all model parameters. It requires the specification of all full conditionals, as described in the Appendix. Within the Bayesian analysis, proper uninformative priors are used. The exact specifications of the priors are given in the example and in the Appendix. Simulated values from the posterior distributions are obtained using the Gibbs sampler. The sampled parameter values can be used to estimate all model parameters, including the attitude of the respondents.

The estimation problem can also be viewed as a missing data problem. The variable of interest, the latent attitudes, cannot be observed directly but usually inferences can be made from observed item responses. However, the observed data consist of randomized item responses. So, the latent attitudes, as well as the true item responses are missing. A popular technique for handling missing data is multiple imputation, see, for example, Rubin (1987, 1996) or Mislevy (1991). Multiple imputations can be used to replace missing values with several potential values from its posterior distribution. In summary, true item responses are sampled given the observed randomized responses, and latent parameter values are sampled given the sampled true item responses.

Simulation

A simulation study was carried out to assess the performance of the Gibbs sampler. Thereafter, results are reported of an example to illustrate the method. The simulated data contained item responses of 1,000 respondents, say students, equally divided over $j = 1, \dots, 20$ groups, say schools. It was assumed that the $k = 1, \dots, 20$ items measure an underlying unidimensional characteristic. The latent variable $\boldsymbol{\theta}$ was generated via a multilevel model, that is,

$$\begin{aligned}\theta_{ij} &= \beta_{0j} + \beta_{1j}X_{ij} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j},\end{aligned}\tag{7}$$

where $e_{ij} \sim \mathcal{N}(0, \sigma_e = 1)$, and $\mathbf{u}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{T})$ with \mathbf{T} a diagonal matrix with elements .1. So, it was assumed that the underlying attitudes of students in different groups were independent. The variable \mathbf{X} can be seen as an individual background variable and was simulated from a normal distribution with standard deviation $\sqrt{.5}$. Via a normal ogive model, Equation 2, true item responses, $\tilde{\mathbf{Y}}$, were simulated, with item discrimination parameters equal to one and difficulty parameters equal to zero. Then, the forced randomized response method was used (see Figure 2) to simulate randomized item response data \mathbf{Y} .

One hundred data sets were generated and analyzed using the Gibbs sampler. All parameter estimates are based on 50,000 iterations and a burn-in period of 5,000 iterations. The convergence of the MCMC chains was checked using the standard convergence diagnostics from the Bayesian output analysis (BOA, <http://www.public-health.uiowa.edu/boa>). The BOA software contains the Geweke convergence test, the Raftery–Lewis test, the Heidelberg–Welch, and the halfwidth test. The diagnostic tests and plots of the sampled values indicated convergence of the MCMC chains. The log-posterior associated with each draw of $(\mathbf{a}, \mathbf{b}, \boldsymbol{\gamma}, \boldsymbol{\sigma}_e, \mathbf{T})$ was computed. Moreover, two parallel MCMC samplers were ran using dispersed initial values. Figure 3 shows sample paths of the log-posterior, the fixed effect γ_{00} , and the Level 1 standard deviation σ_e of the RIRT model, corresponding to one of the simulated data sets. Each individual plot contains the overlaid traces from both parallel chains for a single parameter. The quick mixing of the parameters are obvious, and it can be seen that the Gibbs sampler converged within 5,000 iterations. The corrected Gelman and Rubin convergence diagnostic (Brooks & Gelman, 1998) also suggested good mixing within and between chains. Notice that the mean estimate of the log-posterior values can also be used to compute a Bayes factor.

In this Bayesian analysis, proper but independent noninformative priors were used. For the item parameters, a simultaneous noninformative proper prior was defined ensuring that each item had a positive discrimination index, see the Appendix for specific details. Then, normal priors were assigned to $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ with very large variances to make their means irrelevant. A Wishart prior with small degrees of freedom, with the identity matrix as the precision matrix, was assigned to \mathbf{T}^{-1} . Finally, a gamma prior was assigned to σ_e with small values for the scale and shape parameters.

In this simulation study, the simulated $\boldsymbol{\theta}$ and item parameter values were used to simulate randomized item response data, \mathbf{Y} , and also used to simulate true item response data, $\tilde{\mathbf{Y}}$. The true item response data were analyzed without the randomized response technique. Figures 4 and 5 present the estimated sample posterior mean and 95% credible intervals (CI) of the discrimination and difficulty parameter, respectively, given the true item response data in the upper panel, and given the randomized response data in the lower panel. The actual coverage, that is, the percentage of intervals that contained the true parameter value are presented above each CI. The horizontal dashed line presents the true item parameter values.

The item parameters were estimated given the true item response data and the randomized item response data. It can be seen that the Gibbs sampler gives reasonable results, that is, there is a close agreement between the true parameter values and the estimated means. The item parameter estimates based on the true item response data have smaller credible intervals. Larger posterior variances were computed for the estimates based on the randomized response data because there is less information available from each respondent. When increasing the sample size of the randomized response data the same precision can be obtained. Furthermore, more efficient estimates are obtained when moving p_1 and p_2 further apart from each other. It is remarkable that all estimates of the discrimination parameters are

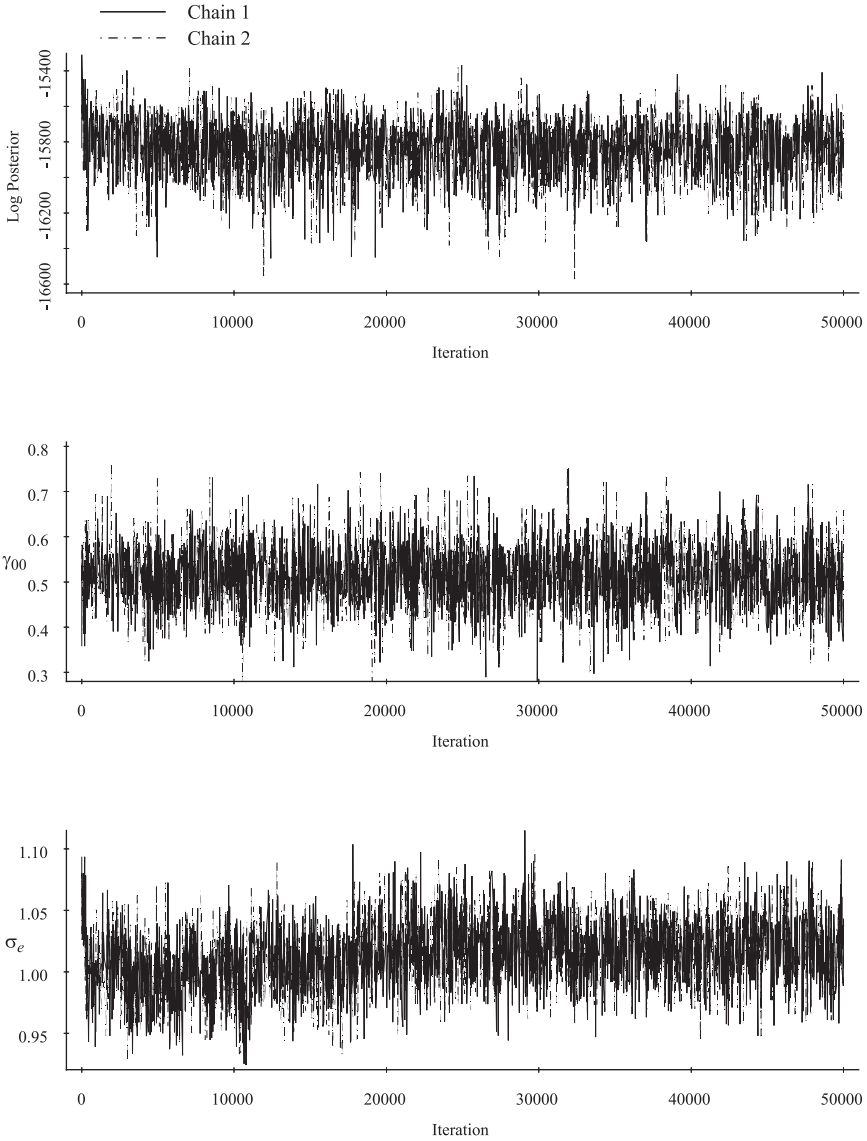


FIGURE 3. Trace plots of the log-posterior, fixed effect γ_{00} , and the Level 1 standard deviation corresponding to the RIRT model.

slightly above one. The model was identified by fixing the scale of the latent variable and this could provoke the small amount of bias in the estimates of the discrimination parameters. To make the outcomes comparable, both models were identified in the same way.

Discrimination Parameter

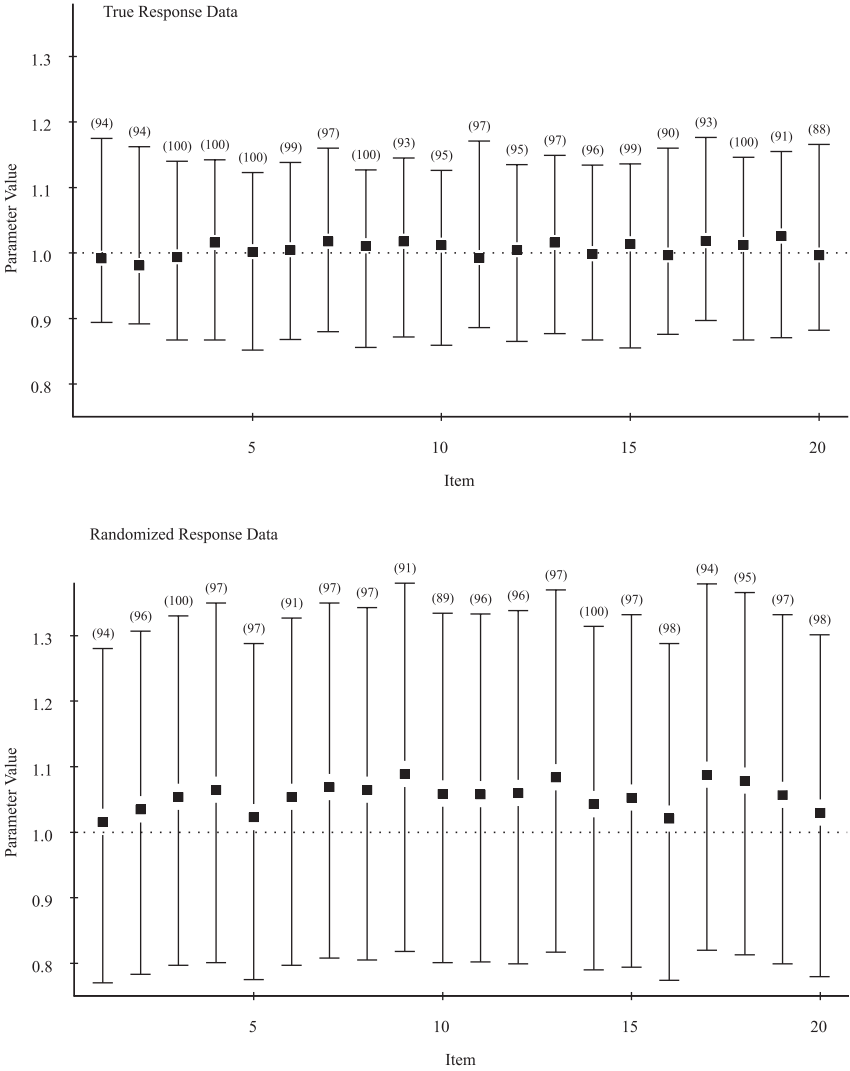


FIGURE 4. Plots of the overall estimates and credible intervals of the discrimination parameters, based on the simulated true item response data and the randomized response data.

Table 1 gives the multilevel parameter estimates. The true generated parameter values are given under the label Gen. The sample estimates of the multilevel parameters given the true latent simulated variable θ are given under the label ML model. These parameter estimates were obtained using a Gibbs sampler for estimating the

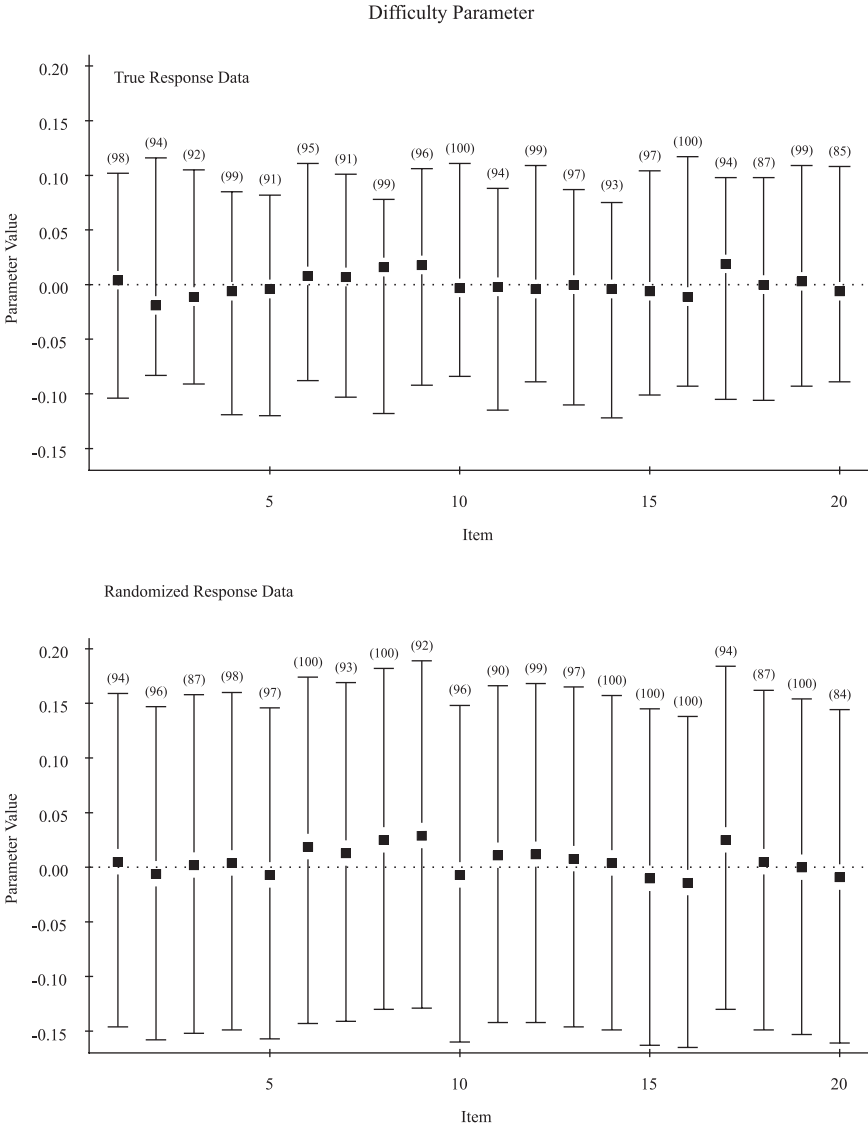


FIGURE 5. Plots of the overall estimates and credible intervals of the difficulty parameters, based on the simulated true item response data and the randomized response data.

parameters of multilevel model (see, e.g., Seltzer, Wong, & Bryk, 1996), with the same kind of prior information. The sample posterior mean, posterior standard deviation, the 95% credible intervals, and the actual coverage, based on the randomized response data, are given under the label RIRT model. Again, the sample posterior

TABLE 1

Results of Simulation Study Regarding Multilevel Model Parameters. Generating Values, Means and Standard Errors of Recovered Values

Fixed	Gen.	ML Model		RIRT Model			
	Coeff.	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	CI	Cov
γ_{00}	0.50	0.51	0.06	0.51	0.08	[0.35, 0.66]	0.96
γ_{10}	0.50	0.51	0.06	0.48	0.05	[0.39, 0.58]	0.94
Random		Var. Comp.	<i>SD</i>	Var. Comp.	<i>SD</i>	CI	Cov
σ_e	1.00	1.00	0.04	1.02	0.02	[0.97, 1.07]	0.98
τ_0	0.32	0.30	0.03	0.33	0.05	[0.21, 0.49]	0.92
τ_{01}	0.00	-0.00	0.02	-0.00	0.03	[-0.06, 0.02]	0.62
τ_1	0.32	0.29	0.04	0.28	0.05	[0.13, 0.38]	0.72

means and standard deviations resemble the true parameter values. A small difference with the true value was obtained in the sample estimate of the parameter τ_1 . It can be stated that the true posterior standard deviation between the slope parameters was underestimated. The sample estimate of this parameter was also underestimated given the true dependent variable. The parameter τ_{01} represents the covariance between the random intercept and random slope parameter. The low coverage percentage of this covariance parameter was caused by the relatively small posterior standard deviations concerning the estimates of τ_{01} . The standard deviations of the parameter estimates regarding the RIRT model are slightly higher in comparison to multilevel model parameter estimates that are based on the true dependent variable. The fixed parameter estimates and the intraclass correlation coefficient are significant. So, the multilevel analysis given the randomized response data, using the MCMC algorithm, did result in the exposure of the fixed and random group effects concerning the latent underlying individual characteristics.

Analyzing Cheating Behavior at a Dutch University

Detecting fraud is difficult, and educational organizations are often not willing to expend the effort required to get to the bottom of cheating cases. On the other hand, student cheating and plagiarism become important problems with the rising possibilities of ways to cheat on exams. The introduction of mobile phones and handheld computers has led to high-tech cheating with web-equipped cell phones or handheld organizers. Today, with the latest developed mobiles, text messages or photographs of test questions are easily sent to others. In 2002, a study was done to assess cheating behavior of students at a University in the Netherlands. The main targets were to investigate the number of students committing fraud, their reasons, and the different ways students cheat on exams.

A sample of studies was drawn from which a stratified sample of students was drawn so that different studies were represented proportional to their total number

of students. The students received an e-mail in which they were asked to cooperate. The forced alternative method was explained in the e-mail to gain the respondents' confidence, so that they were willing to participate and also to answer truthfully. A Web site was developed containing questions concerning cheating on exams and assignments. Each item was a statement, and respondents were asked whether they agreed or disagreed with it. When a student visited the Web site, an on-Web dice server rolled two dice before a question could be answered. The result of both rolls determined if the student was compelled to answer "yes," when the sum of the outcomes were two or three, to say "no" when the sum was twelve, or to answer the sensitive question, in the other cases. In fact, a forced response was automatically given because it is known that some respondents find it difficult to be compelled to lie (see Fox & Tracy, 1986). The forced response technique was implemented with $p_1 = 3/4$ and $p_2 = 2/3$.

The items asked were divided in three groups: investigating ways, frequencies, and reasons to commit fraud. The responses to the set of items contained more information than cheating behavior of the respondents. Therefore, the responses to the items that were assumed to measure the underlying attitude, cheating behavior, were selected. It was investigated, by means of an item analysis, that a subset of 20 items reflected a single underlying characteristic and formed an internally consistent scale. The total number of respondents (698) were divided over $J = 7$ different studies, that is, Computer Science (CS), Educational science and technology (EST), Philosophy of Science (PS), Mechanical Engineering (ME), Public Administration and Technology (PAT), Science and Technology (ST), and Applied Communication Science (ACS).

Frequencies at the item level of cheating can be given from a classical analysis of the RR data (Warner, 1965). In summary, about 25% of the respondents have once cheated on an exam. However, almost 55% of the students admit that they have observed cheating. To obtain more information at the item level and to investigate group mean differences, the RIRT model was estimated given the RR data to the set of 20 items. Various RIRT models were estimated, and each model was identified by fixing the scale of the latent attitude, with mean zero and variance one. Each MCMC procedure contained 50,000 iterations and a burn-in period of 5,000 iterations. Convergence of the MCMC chains was checked using the standard convergence diagnostics from the BOA program. Plots of the runs and the diagnostic tests suggested a convergence of the MCMC chains.

In Figure 6, the top figure presents the estimated posterior distribution of the latent attitudes of the respondents towards cheating. The 10 students with the lowest score on this attitude scale are CS students and the 10 highest scores correspond to six EST, two PAT, and two ACS students. In fact, all students can be ordered with respect to this attitude scale. The bottom figure presents three item characteristic functions (ICF). Each function specifies the probability of a positive response given the value of the latent attitude. The probability of a positive response increases as the level of the latent attitude increases. The three ICFs concern items about ways of cheating. The respondents were asked if they engaged in these

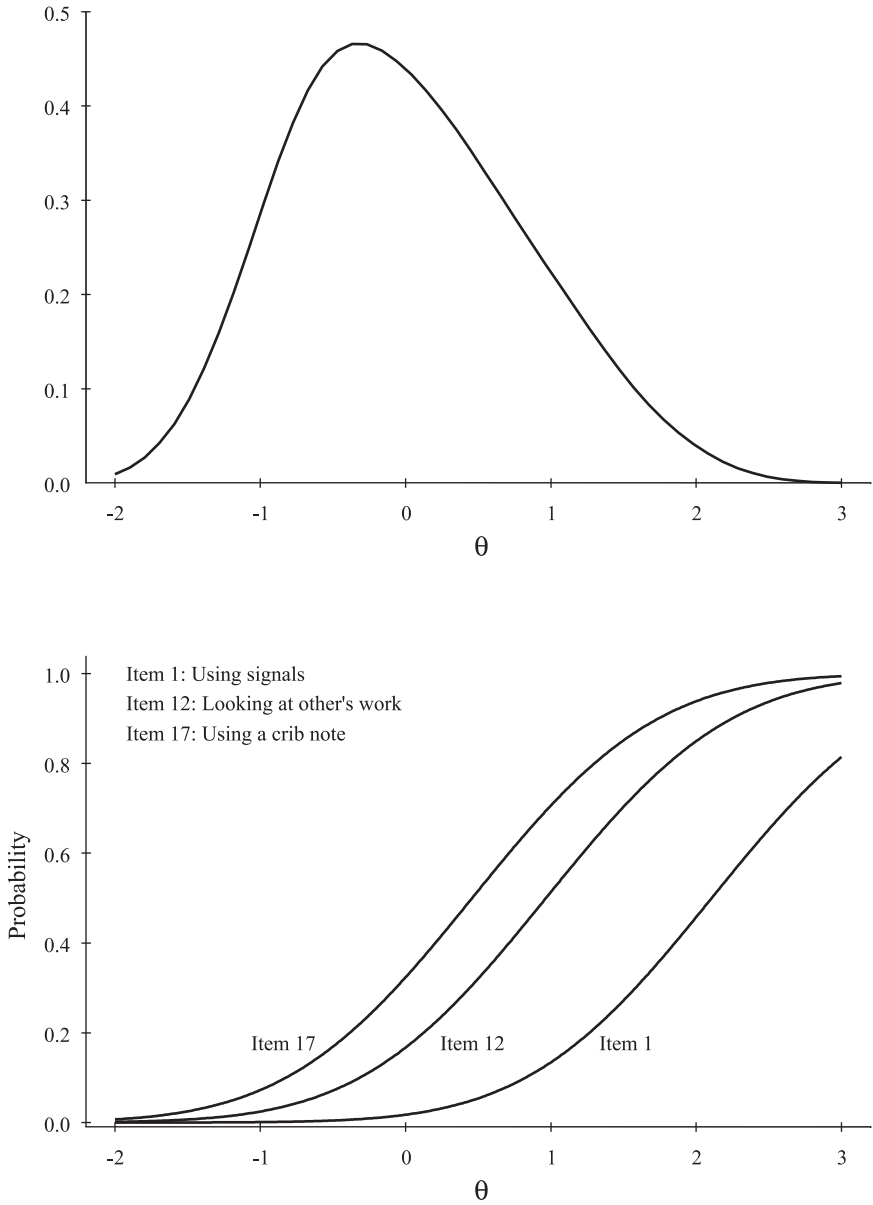


FIGURE 6. Latent attitude distribution and three item characteristic curves about ways of cheating.

specific forms of cheating. Just 5% of the respondents gave a positive response on Item 1, "During an examination, received answers from other students by signaling." Item 12, "Copied an answer by looking at another student's test paper," was positively answered by 24% of the respondents. The ICF of Item 12 shows that, given the overall mean level of the attitude, the probability of a positive response is about 17%. This probability is about 32% for Item 17, "Used forbidden materials, as a crib note, during an exam." So, almost one third of the students have once used a crib note, and this seems to be the most popular method. As a matter of fact, this is in line with other researches regarding cheating (Cizek, 1999).

Differences in Group Means

Interest was focused on investigating differences in group mean attitudes and proportions. First, it was investigated if the mean proportion of positive responses differed across studies. In specific, Item 17, described above, and Item 10, "Written additional information in a reference book that was permitted for answering questions," were considered. The RIRT model was extended with Equation 5 for $k = 10$ and $k = 17$, to investigate if the group mean probabilities of a positive response differed.

In Figure 7, the estimated random group effects for Item 10 and Item 17 are plotted with CIs. For both items, the realized group mean probabilities are given in parenthesis. The labels of the groups include, in parenthesis, the number of students. It is remarkable that far less positive responses were given by Computer Science students. It could be that these students did not trust the method build within a Web site, but no evidence was found to substantiate this. So, the probability of students using a cheat sheet or other forbidden material varies across studies and is particular low for Computer Science students. The intraclass correlation coefficients are .34 and .38, with 95% credible intervals [.16, .52] and [.18, .58], for Items 10 and 17, respectively. So, the fraction of the total variance in the individual response probabilities, attributable to the grouping of students in classes, is quite large for both items.

Another important question was if cheating behavior of the students varied over studies. The differences in attitudes across studies was analyzed with a random effects model. The attitudes of the students were broken down in a contribution from the group, $\gamma_0 + u_j$ and a deviation for each student from the group's contribution, e_{ij} ,

$$\theta_{ij} = \gamma_0 + u_j + e_{ij}, \quad (8)$$

where both residuals, u_j and e_{ij} , are assumed to be independent. Furthermore, u_j is normally distributed with variance σ_u^2 and e_{ij} normally distributed with variance σ_e^2 . Therefore, γ_0 is the general mean, and u_j is the contribution of the group indexed j to the attitudes of the students belonging to it. In Table 2 the parameter estimates are given. Because scaling is the overall mean of the attitudes, γ_0 , zero. The intraclass correlation coefficient is around .17, with a 95% credible interval (.04, .30). This means that around 17% of the total variance, because of

Randomized Item Response Theory Models

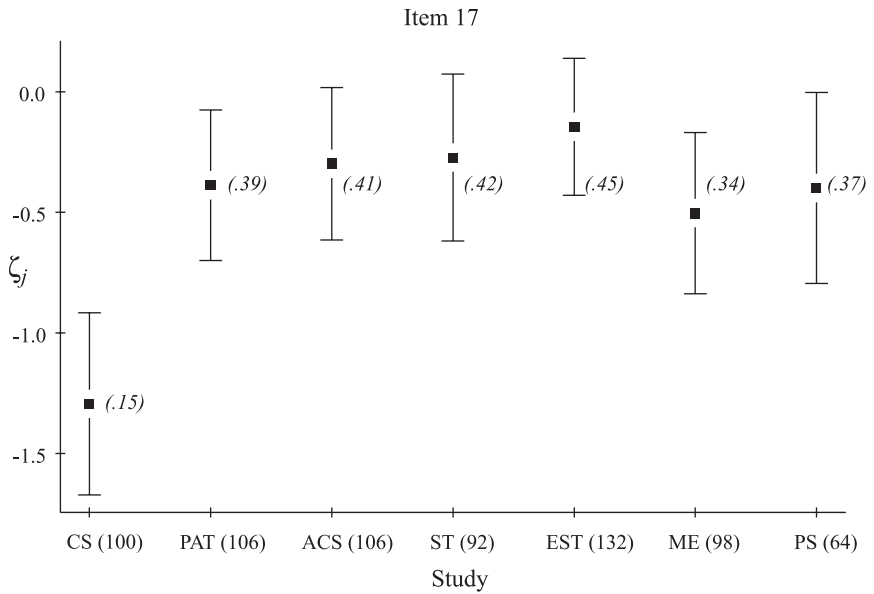
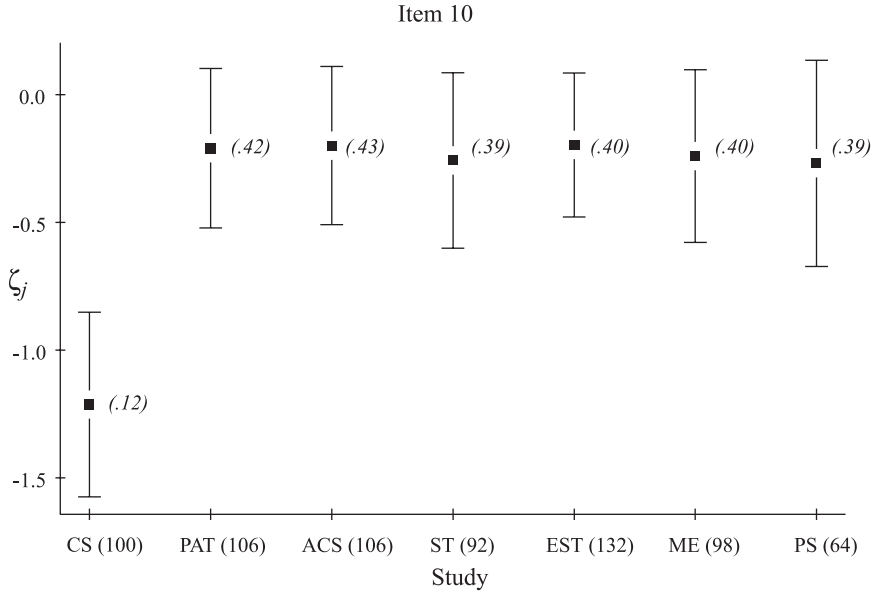


FIGURE 7. Estimates of the random effects, with credible intervals and group mean probabilities, for Items 10 and 17.

TABLE 2
Random Effects ANOVA of Students' Cheating Behavior Nested Within Studies

Fixed Effects	RIRT Model		
	Coefficient	SD	CI
γ_0	-0.02	0.17	[-0.36, 0.32]
Random Effects	Variance Component	SD	CI
σ_e	0.95	0.06	[0.88, 1.00]
σ_u	0.42	0.14	[0.16, 0.70]

individual differences in cheating behavior, can be explained by class differences. The largest difference in attitudes is between the Computer Science and Educational Science and Technology students. The realized group mean attitude, on a standard normal scale, of CS students is $-.73$, and $.27$ for EST students. This reveals itself in different group mean probabilities. For example, the realized estimated probability of a positive response to Item 12 is $.04$ for CS students and $.23$ for EST students.

Studies about students cheating often considered whether male students are more likely to cheat than female students. It is generally believed that female students have a greater tendency to follow rules and fear the consequences should they be caught. On the other hand, female students may have a growing sense that they must cheat to compete with male students, especially in male-dominated classes. The RIRT model, extended with a random effects model, was used to investigate gender effects on cheating behavior taking account for the effect that the students are nested in studies. The random effects model in Equation 8 was extended with explanatory variable *Male*,

$$\theta_{ij} = \gamma_0 + \gamma_1 \text{Male}_{ij} + u_j + e_{ij}, \quad (9)$$

where *Male* was coded as 0 for female and 1 for male students. Table 3 gives the parameter estimates obtained from the MCMC algorithm. The mean attitude of the female students, γ_0 , is zero, because the scale of the latent attitude was fixed with mean 0 and variance 1. The effect of *Male* is slightly negative, indicating that male students are less likely to cheat, but the effect is not significant. It can be concluded that the group mean attitudes toward cheating of male and female students do not differ.

In conclusion, not “everyone’s doing it,” but about 25% of the students admitted they have once cheated on an exam. The analysis with the RIRT model revealed that cheating behavior varied across studies, and that Computer Science students are less likely to cheat. Furthermore, proportion of positive responses to the use of forbidden materials also varied over classes, in specific, EST students are most likely to use forbidden materials.

TABLE 3
*Students' Cheating Behavior Related to Gender with the Intercept Varying
 Between Studies*

Fixed Effects	RIRT Model		
	Coefficient	SD	CI
γ_0	0.01	0.05	[-0.09, 0.01]
γ_1 (Male)	-0.07	0.21	[-0.49, 0.33]
Random Effects	Variance Component	SD	CI
σ_e	0.94	0.06	[0.92, 1.03]
σ_u	0.43	0.15	[0.15, 0.71]

Discussion

Individual attitudes can be estimated when the probability on a positive response, regarding the sensitive question, is modeled by an IRT model. An attitude of an individual toward a sensitive topic can be estimated while obtaining privacy regarding its individual answers. This also makes it possible to compare mean group attitudes and to order respondents on a scale, without knowing the true individual answers. The incorporation of an IRT model opens the possibility for other statistical analyses, such as a one-way classification or multilevel analyses. That is, the IRT model based on RR data can be extended to allow for a nesting of the respondents in groups and explanatory variables concerning characteristics of the respondents or groups. ANOVA could be interesting for comparing proportions of respondents that are cheating conditional on some grouping variable, gender, socioeconomic status or level of education, without asking respondents directly sensitive questions.

All parameters can be estimated simultaneously using the Gibbs sampler. MCMC can be used for computing the true responses, and results in an easy to implement procedure. Further analyses are straightforward, given the sampled parameter values. This in contrast to earlier work on Bayesian inference of randomized response data that results in posterior distributions that are not easy to work with, involves heavy computations, or rely on approximations (see, e.g., Migon & Tachibana, 1997; Winkler & Franklin, 1979). The implementation of the MCMC algorithm can be extended to perform certain model checks. Checking the model assumptions is an important problem. Within the Bayesian framework, posterior predictive checks can provide information regarding the global fit of the model and specific diagnostics can be developed to check such assumptions as local independence, heteroscedasticity, and autocorrelation. Further research will focus on this issue.

The RR method is meant for dichotomous responses but can be extended to polytomous responses using a nominal response model, when there is no a priori ordering of the categories, or a graded response model, when the available

categories can be ordered. Then, using the forced alternative method, the corresponding randomization device should be set up in such a way that with probability p a true response is given, and with probability $1 - p$, a forced response is given in one of the response categories with a certain probability. Respondents may feel that their answers to (sensitive) questions cannot always be captured by true or false. In practice, items with a five-point Likert response format from “1 = strongly disagree” to “5 = strongly agree” are frequently used in educational and psychological measurement.

The posterior variance of the estimated latent characteristic has two components. One is the usual sampling variance associated with the truthful responses. The other is the additional variance caused by the uncertainty associated with the randomized responses. The RR method only works, in the Warner model, when the probability of asking the sensitive question is not equal to $1/2$. The posterior variance of the estimates can be large when this value is close $1/2$ and will be smallest for values near zero or one, provided that the sample size is large enough. On the other hand, to gain the respondents’ confidence, the probability that the sensitive question is chosen cannot be so high as to arouse suspicion. There is a certain trade-off in obtaining honest answers and efficient estimates when using the traditional randomized response technique. The basic idea behind the RR technique is that the respondent must believe that his or her answer cannot be incriminating. However, when applying the RIRT model, inferences can be made at the individual level. So, the subject will probably not cooperate when he or she fully understands this new RIRT model analysis. This paradox complicates the effectiveness of the RIRT model because the method is invented for obtaining truthful answers. Perhaps it is better for researchers to explain only the randomized response model to the respondents to keep them motivated and to keep this new powerful statistical method intact.

Appendix: MCMC Implementation

The parameters are sampled using augmented data. First, nonobserved true item responses are sampled given observed item responses and parameter values. Second, latent continuous true item responses are sampled given the augmented nonobserved true item responses and parameter values. Third, the parameter values are sampled given the augmented continuous true item responses.

Latent Variables

The respondents, indexed ij , respond to the $k = 1, \dots, K$ items. The probability of a positive true response to item k of respondent ij , is modeled by the normal ogive model, and defined in Equation 2. The nonobserved true item responses, $\tilde{\mathbf{y}}$, are sampled given the observations, \mathbf{y} , and values for the model parameters.

- Using Warner’s RR technique: The probability p is known a priori and only affects the precision of the parameter estimates. It can be seen, with the use of Figure 1, that the true nonobserved item responses are Bernoulli distributed. That is,

$$\begin{aligned} \tilde{Y}_{ijk} | Y_{ijk} = 1, \theta_{ij}, a_k, b_k &\sim \mathcal{B}\left(\kappa = \frac{p\pi_{ijk}}{p\pi_{ijk} + (1-p)(1-\pi_{ijk})}\right) \\ \tilde{Y}_{ijk} | Y_{ijk} = 0, \theta_{ij}, a_k, b_k &\sim \mathcal{B}\left(\kappa = \frac{(1-p)\pi_{ijk}}{p(1-\pi_{ijk}) + (1-p)\pi_{ijk}}\right), \end{aligned} \quad (\text{A1})$$

where κ defines the success probability of the Bernoulli distribution and π_{ijk} the probability on a true positive response.

• Using the unrelated-question technique: Here, p_1 , and p_2 are known a priori. Again, the nonobserved true item responses are Bernoulli distributed. From Figure 2 it follows that, after rearranging terms,

$$\begin{aligned} \tilde{Y}_{ijk} | Y_{ijk} = 1, \theta_{ij}, a_k, b_k &\sim \mathcal{B}\left(\kappa = \frac{\pi_{ijk}(p_1 + p_2(1-p_1))}{p_1\pi_{ijk} + p_2(1-p_1)}\right) \\ \tilde{Y}_{ijk} | Y_{ijk} = 0, \theta_{ij}, a_k, b_k &\sim \mathcal{B}\left(\kappa = \frac{\pi_{ijk}(1-p_1)(1-p_2)}{1 - (p_1\pi_{ijk} + p_2(1-p_1))}\right). \end{aligned} \quad (\text{A2})$$

The realizations of the augmented data, $\tilde{\mathbf{y}}$, can be used to sample the normal ogive parameters. To implement the Gibbs sampler for the normal ogive model, Albert (1992) augments the data by introducing independent random variables Z_{ijk} , latent continuous true item responses, which are assumed to be normally distributed with mean $a_k\theta_{ij} - b_k$, and variance equal to one. The augmented observation \tilde{y}_{ijk} can be interpreted as an indicator that the continuous variable with normal density is above or below zero. It follows that

$$Z_{ijk} | \tilde{y}_{ijk}, \theta_{ij}, a_k, b_k \sim \mathcal{N}(a_k\theta_{ij} - b_k, 1), \quad (\text{A3})$$

and $\tilde{y}_{ijk} = I(Z_{ijk} > 0)$.

Parameters

Assuming independence between the item difficulty and discrimination parameter simplifies the choice of the prior because independent sets of parameters may be considered separately. A noninformative prior for the difficulty and discrimination parameter, which ensures that each item will have a positive discrimination index, leads to the simultaneous noninformative proper prior

$$p(\boldsymbol{\xi}) = p(\mathbf{a})p(\mathbf{b}) \propto \prod_{k=1}^K I(a_k > 0) I(a_k, b_k \in \mathbf{A}), \quad (\text{A4})$$

where \mathcal{A} is a sufficiently large bounded interval, in the example $\mathcal{A} = [-100, 100]$. Let \mathbf{Z}_k denote the continuous augmented true item responses to item k of all respondents, and $\boldsymbol{\theta}$ denote all attitude parameters. Let $\boldsymbol{\xi}_k = (a_k, b_k)$, it follows that

$$\xi_k | \mathbf{Z}_k, \boldsymbol{\theta} \sim \mathcal{N}(\hat{\xi}_k, (H' H)^{-1}) I(a_k > 0) I(a_k, b_k \in \mathcal{A}), \quad (\text{A5})$$

where $H = [\boldsymbol{\theta}, -\mathbf{1}]$ and $\hat{\xi}_k$ is the usual least squares estimator following from the linear regression from \mathbf{Z}_k on $\boldsymbol{\theta}$ (Albert, 1992).

The respondents are assumed to be divided over J groups. When explanatory variables are available, the latent attitudes are modeled as a function of Level 1 and Level 2 predictor variables. Here, the full conditional of the latent attitude is described. The full conditional of the respondents' $\boldsymbol{\theta}$ is specified by the linear regression of \mathbf{Z} on $\boldsymbol{\theta}$, with discrimination parameters \mathbf{a} as the regression coefficients, and the linear regression of $\boldsymbol{\theta}$ on the Level 1 characteristics, with random regression coefficients $\boldsymbol{\beta}$ and variance parameter σ_e^2 . The attitude parameters are distributed as a mixture of normal distributions, and the full conditional is again normally distributed. Let \mathbf{X} contain the Level 1 characteristics, it follows that

$$\theta_{ij} | \mathbf{Z}_{ij}, \xi_k, \boldsymbol{\beta}_j, \sigma^2 \sim \mathcal{N}\left(\frac{\hat{\theta}_{ij}/v + \mathbf{X}_{ij}\boldsymbol{\beta}_j/\sigma_e^2}{v^{-1} + \sigma_e^{-2}}, \frac{1}{v^{-1} + \sigma_e^{-2}}\right), \quad (\text{A6})$$

where

$$v = \left(\sum_{k=1}^K a_k^2\right)^{-1}$$

and $\hat{\theta}_{ij}$ the least squares estimator following from the regression of $\mathbf{Z}_{ij} + \mathbf{b}$ on \mathbf{a} . The posterior mean of Equation (A6) has the form of a shrinkage estimator, as the sampling variance v of $\hat{\theta}_{ij}$ increases, the relatively weight of $\mathbf{X}_{ij}\boldsymbol{\beta}_j$ increases. This means that informative Level 1 characteristics increase the precision of the attitude estimates. The full conditionals of the random effect or the multilevel model parameters are described in, among others, Fox and Glas (2001), Seltzer, et al. (1996), and Zeger and Karim (1991). In this article, proper noninformative priors were used. A prior for the variance at Level 1, σ_e^2 , was specified in the form of an inverse-gamma (IG) distribution with shape and scale parameters, $(n_0/2, n_0 S_0/2)$, and $n_0 = 0.0001$, $S_0 = 1$. An inverse Wishart prior distribution for the variance parameter at Level 2 with small degrees of freedom, but greater than the dimension of $\boldsymbol{\beta}_j$. A normal distributed prior for the fixed effects with a large variance parameter.

Consider Equation 5, where it is of interest to test the group mean proportions of positive responses per item. With the introduction of the augmented variables \mathbf{Z} , Equation 5 can be written as a random effects model,

$$Z_{ijk} = \mu_k + \zeta_{jk} + \epsilon_{ijk} \quad (\text{A7})$$

with $\zeta_{jk} \sim \mathcal{N}(0, \sigma_\zeta^2)$, and $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma_\epsilon^2)$. From Equation A7, it follows that ζ_{jk} , with n_j respondents in group j , is normally distributed with parameters,

$$E(\zeta_{jk} | \mathbf{z}, \sigma_\epsilon^2, \sigma_\zeta^2, \mu_k) = \frac{n_j \sigma_\zeta^2}{\sigma_\epsilon^2 + n_j \sigma_\zeta^2} (\bar{z}_{jk} - \mu_k)$$

$$\text{Var}(\zeta_{jk} | \mathbf{z}, \sigma_\epsilon^2, \sigma_\zeta^2) = \frac{\sigma_\epsilon^2 \sigma_\zeta^2}{\sigma_\epsilon^2 + n_j \sigma_\zeta^2},$$

where \bar{z}_{jk} is the group mean of the individual augmented z_{ijk} values. Inverse-gamma prior distributions with $n_0 = .0001$, $S_0 = 1$, were specified for the variance parameters, σ_ϵ^2 and σ_ζ^2 . The full conditional of the random effects regression parameters, μ_k , σ_ϵ^2 and σ_ζ^2 , can be found in Gelman, Carlin, Stern, and Rubin (2004, pp. 390–392).

References

- Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17*, 251–269.
- Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational & Graphical Statistics, 7*, 434–455.
- Cizek, G. J. (1999). *Cheating on tests. How to do it, detect it, and prevent it*. Mahwah, NJ: Erlbaum.
- Fox, J. A., & Tracy, P. E. (1986). *Randomized response*. Beverly Hills, CA: Sage.
- Fox, J.-P. (2004). Modelling response error in school effectiveness research. *Statistica Neerlandica, 58*, 138–160.
- Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*, 269–286.
- Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85*, 398–409.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). London: Chapman & Hall.
- Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*, 721–741.
- Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.) (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall.
- Greenberg, B. G., Abul-Ela, A., Simmons, W. R., & Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association, 64*, 520–539.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Migon, H. S., & Tachibana, V. M. (1997). Bayesian approximations in randomized response model. *Computational Statistics & Data Analysis, 24*, 401–409.
- Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177–196.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
- Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91*, 473–489.
- Scheers, N. J., & Dayton, C. (1988). Covariate randomized response model. *Journal of the American Statistical Association, 83*, 969–974.

- Seltzer, M. H., Wong, W. H., & Bryk, A. S. (1996). Bayesian analysis in applications of hierarchical models: Issues and methods. *Journal of Educational and Behavioral Statistics, 21*, 131–167.
- Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association, 82*, 528–550.
- Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 60*, 63–69.
- Winkler, R. L., & Franklin, L. A. (1979). Warner's randomized response model: A Bayesian approach. *Journal of the American Statistical Association, 74*, 207–214.
- Zeger, S. L., & Karim, M. R. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association, 86*, 79–86.

Author

J.-P. FOX; Fox@edte.utwente.nl.

Manuscript received February 3, 2004

Revision received April 19, 2004

Accepted April 20, 2004

Fox—Author Queries

1. AU: “behaviors” (plural?)

2. AU: 1995 correct? If not, please add Warner (1994) to Refs.

3. AU: Sense? Please recast sentence.

4. AU: In Eq. A6, lower case “vee” or l.c. Greek nu?

5. AU: Please use complete sentence.

6. AU: Pls supply complete author information, including title, affiliation, mailing address, and areas of interest.