Using Item Response Theory to Obtain Individual Information From Randomized Response Data: An Application Using Cheating Data

Jean-Paul Fox and Rob R. Meijer

Applied Psychological Measurement 2008; 32; 595 originally published online Apr 16, 2008;
DOI: 10.1177/0146621607312277

Using Item Response Theory to Obtain Individual Information From Randomized Response Data: An Application Using Cheating Data

Jean-Paul Fox and Rob R. Meijer, University of Twente

The authors discuss a new method that combines the randomized response technique with item response theory. This method allows the researcher to obtain information at the individual person level without knowing the true responses. With this new method, it is possible to compare groups of individuals by means of analysis of variance or regression analysis. To illustrate the advantages of this new method, 349 students of a university in the Netherlands were surveyed with respect to frequency and reasons to cheat on exams, and students' attitudes toward cheating were investigated. Thirty-six items tapping different types of cheating behavior were used to measure attitude toward cheating, and responses to questions were obtained via a randomized response technique. The new method was used to explain differences among students' attitudes using different kinds of background information. *Index terms: ANOVA, attitudes, cheating, item response theory, randomized response data*

In psychological and educational assessment, people often do not respond truthfully when asked personal or sensitive questions. 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. Any research study that uses self-report measures runs the risk of response bias. A well-known example is that persons respond in socially desirable ways (Fowler, 2002).

Warner (1965) developed a data collection procedure, the randomized response (RR) technique, that allows researchers to obtain sensitive information while guaranteeing privacy to respondents. This method encourages greater cooperation from respondents and reduces their motivation to falsely report their attitudes. In the traditional RR technique, respondents are confronted with two mutually exclusive choices: "I belong to Group A," and "I do not belong to Group A." A randomization mechanism is used to choose between the two statements (e.g., tossing a die or using a spinner). The randomization is performed by the interviewee, and the interviewer is not permitted to observe the outcome of the randomization. The interviewee responds to the question selected by the randomization device, and the interviewer knows only the response. The respondent's privacy or anonymity is well protected because no one but the respondent knows which question was answered. It is assumed that respondents are more willing to provide honest answers with this technique because their answers do not reveal any information about themselves.

The RR technique was developed in an attempt to improve the quality of self-reported survey research, but it is not very often applied in an educational or psychological context. Numerous empirical studies have shown that the RR technique results in higher estimates of sensitive characteristics when

Applied Psychological Measurement, Vol. 32 No. 8, November 2008, 595–610
DOI: 10.1177/0146621607312277

compared with direct questioning (e.g., see Scheers & Dayton, 1987; van der Heijden, van Gils, Bouts, & Hox, 2000). Scheers and Dayton (1987) found that underreporting of five academic cheating behaviors ranged from 39% to 83% when responses to an anonymous questionnaire were compared to estimates using an RR technique. They concluded that estimates based on anonymous questionnaires may result in severe underestimation of sensitive behaviors. In a psychological context, Donovan, Dwight, and Hurtz (2003) used the RR technique to estimate the base rate of entry-level job applicant faking during the application process. Their results revealed that a substantial number of recent job applicants reported engaging in varying degrees of misrepresentation and that the base rate for faking was strongly related to both the severity and verifiability of the deceptive behavior.

The traditional RR technique enables the researcher to compute the proportion of the population engaging in a particular kind of behavior or, in general, belonging to Group A. However, further analysis of the RR data is limited because the individual item responses, denoted as true item responses, are randomized before observing them and are therefore unknown. For example, it is not possible to interpret individual response patterns directly (due to observing RRs) or to compare individuals or groups of individuals by means of analysis of variance (ANOVA) or regression analysis because the within-population differences corresponding to the proportion of the population engaging in a particular kind of behavior are unknown. In the present study, the RR technique was combined with an item response theory (IRT; Lord & Novick, 1968) model that enabled an analysis at the individual person level. A major advantage of using an IRT model is that item characteristics can be explored and related to individual differences in the trait or attitude being measured. Thus, by means of a combined use of RR data and IRT, more valid and reliable information can be obtained without losing information at the individual person level. A simulation study showed that person parameter estimates given randomized item response data were close to the generated values, and person parameter estimates given true item response data led to comparable differences.

This study is organized as follows. First, the basic principles of the RR technique and the common traditional method of analyzing RR data are introduced. The next section illustrates how the RR technique and IRT can be combined. Finally, the combined use of the RR technique and IRT is illustrated with data from student cheating.

The RR Technique

A commonly used RR technique is the unrelated question design. This design, described by Horvitz, Shah, and Simmons (1967), involves pairs of questions, one innocuous and one concerned with the sensitive behavior being investigated. In the unrelated question design, the second question is not related and is completely innocuous, and the probability of a positive response is known. The unrelated question can also be built into the randomizing device. In this forced alternative method, the randomization device does not select the unrelated innocuous question but directly generates a random (forced) response. So a randomization device determines if the respondent is forced to answer positively, negatively, or truthfully to the sensitive question. The researcher does not know which statement is selected by the randomization device. The aim is to estimate the true positive proportion of responses to the sensitive question and related confidence intervals in the population. This can be done as follows.

Let Y denote the dichotomous observed response (true/false or yes/no), and let $P(Y = 1)$ denote the probability of an observed positive response from a respondent in the population. The probability of an observed positive response can be linked to the probability of a positive response to the sensitive question via the linear equation

$$P(Y = 1) = p_1\pi + (1 - p_1)p_2, \quad (1)$$

where p_1 is the probability that a respondent has to answer the sensitive question and p_2 the probability that a respondent has to give a forced positive response. Both probabilities are defined by the randomization device. Parameter π denotes the probability of obtaining a positive response to the sensitive question. In a specific population, π is the proportion of respondents answering positively to the sensitive question.

Binomial Model for RR Data

There is a standard binomial approach for analyzing univariate RR data. This approach can be followed for each of the item responses; however, dependencies among individuals' item responses are neglected.

In the real data example, interest is focused on individual and group differences in attitudes toward cheating. Therefore, the observed RRs are assumed to be nested within groups. More generally, the observed RRs of individuals $i\,(i=1,\ldots,n_j)$ in group $j\,(j=1,\ldots,J)$ to item $k\,(k=1,\ldots,K)$, $\mathbf{Y}_{jk}$, are assumed to be binomially distributed; that is,

$$Y_{ijk} \sim B(n_j, p_1\pi_{jk} + (1-p_1)p_2). \tag{2}$$

with success probability $p_1\pi_{jk} + (1-p_1)p_2$, where π_{jk} is the group-specific true proportion of positive responses. Let $\bar{y}_{jk} = \sum y_{ijk}/n_j$ denote the mean RR to item k of all individuals in group j, say, n_j, respectively. Then, via equation (1) and well-known properties of a binomial random variable, an estimate of the group-specific true proportion of positive responses can be derived (e.g., see Horvitz et al., 1967):

$$\begin{array}{rcl} \bar{y}_{jk} & = & p_1\hat{\pi}_{jk} + (1-p_1)p_2, \\ \hat{\pi}_{jk} & = & \big(\bar{y}_{jk} - (1-p_1)p_2\big)\big/p_1. \end{array} \tag{3}$$

In the same way, the corresponding estimated variance equals

$$\hat{V}(\hat{\pi}_{jk}) = \bar{y}_{jk}\big(1-\bar{y}_{jk}\big)\big/n_j p_1^2. \tag{4}$$

In this binomial modeling approach, it is assumed that the individuals' responses to item k are independent. Note that the estimates cannot be directly linked to individual explanatory variables because they are defined at a group level. The difference between two group proportions $(j=1,2)$ can be tested using the statistic

$$z = \frac{\hat{\pi}_{1k} - \hat{\pi}_{2k}}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}, \tag{5}$$

where $\hat{p}$ is the combined proportion of positive responses over groups,

$$\hat{p} = \frac{\sum_i y_{i1k} + \sum_i y_{i2k}}{n_1 + n_2}, \tag{6}$$

and for large n_1 and n_2 the sampling distribution of z is standard normal. When comparing differences among more than two groups, proportions can be tested using a chi-square statistic. These statistics can be used to compare true proportions of positive responses across groups. Note that the analyses are limited to the item level and cannot be directly extended to observations from multiple items. Furthermore, as pointed out by a reviewer, the statistic has an asymptotic normal sampling distribution. However, for extreme response probabilities, the normality (chi-square) assumption holds only for a very large number of observations. For these and previously stated reasons, an alternative method is proposed for testing true proportions of positive responses.

An IRT Model for RR Data

The combination of IRT with the RR technique results in a new method to measure traits or attitudes, denoted as θ, without knowing the true individual answers. To explain individual differences in the trait being measured, an IRT model is assumed for the true item responses. A relationship is specified between the observed RR data and the true item response data. Assume, for the moment, that the true individual item responses are known and denoted as $\tilde{\mathbf{Y}}$. The probability of a positive true response for a respondent, indexed i, to item k is defined by the two-parameter normal ogive model; that is,

$$\pi_{ik} = P\left(\tilde{Y}_{ik} = 1|\theta_i, a_k, b_k\right) = \Phi(a_k\theta_i - b_k), \tag{7}$$

where Φ (.) denotes the cumulative normal distribution function, and a_k and b_k are the item's discrimination and difficulty parameters, respectively. The probability of observing a positive RR can be linked to the probability of observing a true item response. This is done by substituting equation (7) into equation (1); that is,

$$\begin{aligned} P(Y_{ik} = 1|\theta_i, a_k, b_k) &= p_1 P\left(\tilde{Y}_{ik} = 1|\theta_i, a_k, b_k\right) + (1 - p_1)p_2 \\ &= p_1\Phi(a_k\theta_i - b_k) + (1 - p_1)p_2 \\ &= p_1\pi_{ik} + (1 - p_1)p_2, \end{aligned} \tag{8}$$

where probabilities p_1 and p_2 of the randomization device are known a priori. The RR sampling design is such that with probability p_1 a true response is observed and with probability $1 - p_1$ a forced response is observed.

The randomized item response data cannot be analyzed directly due to the fact that for each respondent some of the observations are forced responses. The observed randomized item responses are connected with the true item responses through the RR technique. With this mechanism, probability statements can be made about the true item responses given the observed randomized item responses. Another variable, Z, is defined to equal 1 when the randomization device determines that the item has to be answered truthfully and to equal 0 otherwise. Now, the probability that a true positive response to item k is observed given a positive RR can be derived:

$$\begin{aligned} P\left(\tilde{Y}_{ik} = 1|Y_{ik} = 1\right) &= \frac{\sum_{z=0,1} P\left(\tilde{Y}_{ik} = 1, Y_{ik} = 1|Z_{ik} = z\right)P(Z_{ik} = z)}{\sum_{z=0,1} P(Y_{ik} = 1|Z_{ik} = z)P(Z_{ik} = z)} \\ &= \frac{\pi_{ik}p_1 + \pi_{ik}p_2(1 - p_1)}{\pi_{ik}p_1 + p_2(1 - p_1)}, \end{aligned} \tag{9}$$

where the denominator follows from equation (8). Equation (9) displays the conditional probability of a true positive response given an observed positive RR. In this way, the true item responses can be associated with the observed randomized item responses.

Parameter Estimation

The log likelihood of the randomized item response model given RR data resembles the usual log likelihood of an IRT model nested within an RR sampling design; that is,

$$\log L(\mathbf{a}, \mathbf{b}, \boldsymbol{\theta}|\mathbf{y}) = \sum_i \sum_k y_{ik}\log\varphi_{ik} + (1 - y_{ik})\log(1 - \varphi_{ik}), \tag{10}$$

where φ_{ik} is given by equation (8) (i.e., $\varphi_{ik} = p_1\Phi(a_k\theta_i - b_k) + (1 - p_1)p_2$), for $k = 1, \ldots, K$ items and $i = 1, \ldots, N$ persons. Although the parameters p_1 and p_2 of the RR model are known, the usual

formulas for estimating the normal ogive model parameters, first- and second-order derivatives of a normal ogive log-likelihood function (e.g., see Baker, 1992), differ from the first- and second-order derivatives of the log-likelihood function of equation (10). This is engendered by the fact that the probability of a positive response is influenced by the RR model with parameters p_1 and p_2. It follows that in the estimation of the randomized item response model parameters, the RR model has to be taken into account, and a standard IRT software package such as BILOG (Mislevy & Bock, 1990) cannot be used.

The model parameters can be estimated simultaneously using Markov chain Monte Carlo (MCMC) methods. This is a simulation-based approach for computing the posterior distributions of the model parameters. Albert (1992) and Patz and Junker (1999a, 1999b) described MCMC implementations for simulating draws from the conditional posterior distributions of the item parameters and the latent attitude parameter given the item response data for a normal ogive or a logistic IRT model, respectively. If the true item response data were observed, random draws were obtained from the corresponding conditional distributions of the model parameters. The key idea is that given the true item response data, all other model parameters are easily derived. The true item responses are of course unknown, but the distribution of the true item responses given the observed randomized item responses is known; see equation (9).

Step 1. The true item response data, $\tilde{\mathbf{Y}}$, can be sampled given the observed RRs' $\mathbf{Y}$:

$$\begin{aligned} \tilde{Y}_{ik}|Y_{ik} &= 1, \theta_i, a_k, b_k \sim B\left(\frac{\pi_{ik}(p_1 + p_2(1-p_1))}{\pi_{ik}p_1 + p_2(1-p_2)}\right), \\ \tilde{Y}_{ik}|Y_{ik} &= 0, \theta_i, a_k, b_k \sim B\left(\frac{\pi_{ik}(1-p_1)(1-p_2)}{1-(\pi_{ik}\,p_1 + p_2(1-p_2))}\right), \end{aligned} \tag{11}$$

for $k = 1, \ldots, K$ and $i = 1, \ldots, N$, where $B(\nu)$ denotes a Bernoulli distribution with success probability ν. Given the true item response data, an MCMC scheme for simulating draws from the conditional distribution of the IRT model parameters can be applied.

Step 2. It follows that draws can be obtained from the following distributions:

1. $p(a_k|b_k, \theta_i, \tilde{y}_{ik})$,
2. $p(b_k|a_k, \theta_i, \tilde{y}_{ik})$,
3. $p(\theta_i|a_k, b_k, \tilde{y}_{ik})$.

The conditional posterior distributions are fully specified in Albert (1992) and Patz and Junker (1999a, 1999b), among others. In conclusion, first, true item responses are simulated given the RR data; second, all model parameters are simulated given the true item response data. This procedure is iterated until enough samples are drawn to obtain accurate parameter estimates; see J.-P. Fox (2005) for a detailed discussion. The software for the estimation procedure can be obtained from the authors and is also available on the Internet. The program consists of a set of Fortran routines that can be applied in the statistical package R. Different models can be compared using the Bayesian information criterion (Raftery, 1995) statistic that can be obtained via the estimated marginal likelihood from the program.

Simulation Study

This simulation study showed that the MCMC person parameter estimates were close to the true simulated person parameter values in the cases of true item response data and randomized item response data. Therefore, difficulty parameter values were generated from a standard normal

distribution, and discrimination parameter values were generated from a normal distribution with a mean of 1 and a variance of .2. Person parameter values were generated from a standard normal distribution. True item response data were generated with a success probability according to equation (7) given this set of simulated parameter values. Randomized item response data were generated with a success probability according to equation (8), $p_1 = 4/5$ and $p_2 = 2/3$, and the same set of parameter values that were used for generating the true item response data. The randomized item response data were analyzed using the described MCMC procedure. The true item response data were analyzed using the MCMC procedure without the extra sampling step in equation (11) for handling RR data.

For two conditions, the average absolute difference between the true and estimated person parameters was computed given true response data and RR data. For 1,000 persons and 20 items, this difference equaled .381 given true response data and .385 given RR data. For 500 persons and 10 items, the difference equaled .421 given true response data and .428 given RR data. It was concluded that the estimated person parameters were in both cases comparably close to the true values.

Testing for Differences Between Groups

The normal population distribution of the attitude parameter can be extended to account for a grouping structure. Let the grouping structure be represented by indicator variables that take on the values 1, -1, or 0, and let these indicator variables be stored in a design matrix $\mathbf{x}$. It follows that the population distribution of the attitude parameters is normally distributed with variance σ^2 and mean $\mathbf{x}\boldsymbol{\lambda}$, where $\boldsymbol{\lambda}$ represents the factor effects in this Bayesian ANOVA study. This model is related to the mixture IRT model that accommodates latent group membership into an IRT model (e.g., see Rost, 1990, 1991) and can be used to investigate individual differences given item response data. Note that in this case the clustering of individuals is known. Furthermore, the attitude estimates corresponding to the randomized IRT model with covariate information are expected to have smaller variances than those corresponding to a randomized IRT model without covariate information.

Step 3. The steps of the MCMC algorithm for sampling the model parameters can be extended with draws from the full conditional posterior distributions of parameters σ^2 and $\boldsymbol{\lambda}$. That is, draws can be obtained from the following distributions:

1. $p(\boldsymbol{\lambda}|\boldsymbol{\theta}, \sigma^2)$,
2. $p(\sigma^2|\boldsymbol{\theta}, \boldsymbol{\lambda})$.

The conditional posterior distributions of Step 3 can be found in, for example, Gelman, Carlin, Stern, and Rubin (1995, chap. 8). In this approach, the conditional posterior distribution of θ changes slightly because the extended population distribution of θ should be taken into account in deriving the full conditional posterior distribution (e.g., see J.-P. Fox & Glas, 2001).

Attention is focused on a point null hypothesis of the form $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0$ and an alternative hypothesis $\boldsymbol{\lambda} \neq \boldsymbol{\lambda}_0$. A Bayesian test of significance is performed on this null hypothesis at a level of significance of α. This is done by constructing a confidence interval from the posterior distribution and rejecting the null hypothesis if and only if $\boldsymbol{\lambda}_0$ is outside this interval. Note that the null hypothesis $\boldsymbol{\lambda}_0 = 0$ is of particular interest because this indicates whether the mean attitudes toward cheating are the same for all groups. A Bayesian p value can be defined as 1 minus the content of the confidence interval that just covers $\boldsymbol{\lambda}_0$, which equals

$$P[p(\boldsymbol{\lambda}|\mathbf{y}) \leq p(\boldsymbol{\lambda}_0|\mathbf{y})|\mathbf{y}]. \tag{12}$$

The null hypothesis is rejected when the Bayesian p value is less than or equal to a significance level α. For the linear model, Box and Tiao (1973, chap. 2) have shown that equation (12) is equal to the quantity

$$P\left[F_{k,\nu} \leq \frac{(\boldsymbol{\lambda}_0 - \boldsymbol{\lambda})^t \mathbf{x}^t \mathbf{x} (\boldsymbol{\lambda}_0 - \boldsymbol{\lambda})}{k s^2} \middle| \mathbf{y}\right], \tag{13}$$

where $F_{k,\nu}$ is an F variable with k (and rank of $\mathbf{x}$), $\nu = N - k$, and s^2 is the observed sampling variance. The computation of the Bayesian p value in equation (13) can be done by computing this probability in each iteration of the MCMC algorithm given the sampled values of the parameters after the burn-in period. Eventually, the average of all computed values is considered to be an estimate of the Bayesian p value. This means that Bayesian ANOVA tests can be performed within the estimation procedure, and it is not necessary to rely on parameter estimates. This strategy is easily extended to multifactor studies. In the same way, let $\boldsymbol{\lambda}_s$ denote an s-dimensional subset of the k-dimensional parameter $\boldsymbol{\lambda}$ with covariance matrix $\mathbf{x}_s$. Box and Tiao (1973, chap. 2) showed that a comparable F test such as the one in equation (13) can be derived for an s-dimensional subset of parameter $\boldsymbol{\lambda}$.

At the item level, a comparable procedure can be applied to test whether the mean response probabilities to an item are the same for all groups. Let a design matrix $\mathbf{x}$ contain indicator variables that represent a certain grouping structure. Let π_{ijk} denote the probability of a positive response to item k of individual i in the jth level ($j = 1, \ldots, J$) of a factor. This can be easily generalized to multiple factors. A probit link function was used to transform these probabilities to a new dependent variable that has values on the real line (McCullagh & Nelder, 1989). It is assumed that

$$\Phi^{-1}(\pi_{ijk}) \sim N\left(\mathbf{x}\boldsymbol{\eta}, \tau^2\right), \tag{14}$$

where Φ^{-1} is the inverse cumulative normal distribution function and $\boldsymbol{\eta}$ represents the factor effects in this Bayesian ANOVA study. The full conditional posterior distribution of parameter $\boldsymbol{\eta}$ and τ^2 follow in the same way as in the Bayesian ANOVA study for $\boldsymbol{\theta}$.

Step 4. Draws can be obtained from

1. $p(\boldsymbol{\eta} | \boldsymbol{\pi}, \tau^2)$,
2. $p(\tau^2 | \boldsymbol{\pi}, \boldsymbol{\eta})$.

Note that these additional draws can be made within the MCMC algorithm for estimating all other model parameters. That is, this step is performed after Steps 1-3. Before performing Step 4, $\boldsymbol{\pi}$ is computed given the sampled item parameter values. Interest is focused on the null hypothesis $\boldsymbol{\eta} = 0$. The inverted individual probabilities of a positive response corresponding to a specific item can be used to test whether the group means of response probabilities corresponding to positive responses are the same for all groups. Finally, a Bayesian p value can be defined in the same way as the one in equation (13) and can be used to test the null hypothesis. This test statistic is computed via the MCMC algorithm. Note that multifactor studies can be performed in the same way by defining a design matrix and testing whether different subsets of $\boldsymbol{\eta}$ are equal to zero. This test on proportions is known to be sensitive to the assumption of normality, and if $\boldsymbol{\pi}$ is small or large, the population distribution of the transformed probabilities can be very skewed. In that case, it takes a fair number of samples to achieve symmetry.

A nonparametric alternative to the parametric approach in equation (14) is the Kruskal-Wallis test. This test uses only the ordering of the group means of response probabilities. In each iteration of the MCMC algorithm, all N response probabilities are pooled, arranged in order of size, and

ranked. Ties are assigned the same rank. After regrouping the response probabilities, the sum of the ranks in each group is computed. Subsequently, the Kruskal-Wallis test is computed in each MCMC iteration, and the mean value of all computed values is considered to be an estimate of the test statistic. For reasonably large group sizes (five or more) the Kruskal-Wallis test statistic is approximately chi-square distributed with $J-1$ degrees of freedom.

An Application Using Cheating Data

Participants and Sampling Method

The usefulness of a combined IRT model and RR data was illustrated in the context of student cheating. Ample evidence that cheating occurs in college has been provided by research results from the United States and the United Kingdom (e.g., see Anderman & Midgley, 2004; Cizek, 1999; Davis, Grover, Becker, & McGregor, 1992; Murdock, Miller, & Kohlhard, 2004; Newstead, Franklyn-Stokes, & Armstead, 1996; Whitley, 1998). The present study focused on academic cheating, knowing that most students are not eager to share information about their practices and attitudes toward cheating. It is clear that self-report measures can be easily faked and that faking may reduce the utility of these measures.

Data were available from 349 students from a university in the Netherlands. There were 229 male students and 120 female students from one of seven main disciplines at this university: computer science (CS), educational science and technology (EST), philosophy of science, mechanical engineering, public administration and technology, science and technology, and applied communication sciences. Within these seven disciplines, a stratified sample of students was drawn such that different majors were represented proportional to their total number of students.

Procedure and Measure

The students received an e-mail in which they were asked to participate in the survey. The forced alternative method was explained to increase the likelihood that students would (a) participate in the study and (b) answer the questions truthfully. A Web site was developed containing 36 statements concerning cheating on exams and assignments. Students were asked whether they agreed or disagreed with the statement. When a student visited the Web site, an on-Web dice server rolled two dice before a question could be answered. The student answered ''yes'' when the sum of the outcomes equaled 2, 3, or 4; answered ''no'' when the sum equaled 11 or 12; or answered the sensitive question truthfully. In reality, a forced response was automatically given because it is known that some respondents find it difficult to lie (e.g., J. A. Fox & Tracy, 1986). The forced response technique was implemented with $p_1 = 3/4$ and $p_2 = 2/3$.

There were 36 items administered concerning types of cheating. Types of cheating or academic dishonesty were measured by different items to capture the whole range of cheating. The list of 36 items contained many items from the list used by Newstead et al. (1996). IRT fit analyses showed that the 36 types-of-cheating items constituted a fairly unidimensional scale (for the content of these items, see Table 2). Detailed information about the fit analyses can be obtained from the authors.

In the cheating literature, several variables such as personal attributes, background characteristics, or situational factors have been shown to be related to cheating behavior. For example, earlier cheating research focused on estimating population proportions of cheating across locations (small/large university), school types (private/public), or individual differences (male/female; e.g., see Davis et al., 1992). In the present study, background information was collected with respect to age, gender, year of major, number of hours per week spent on the major (less than 10 hours, 10-20 hours,

20-30 hours, 30-40 hours, or more than 40 hours), place of living (on campus, in the city, or with parents), and lifestyle (active or passive). Respondents were guaranteed confidentiality, and the questionnaires were filled in anonymously.

Results

Analysis of differences in attitudes. The individual attitudes toward cheating, denoted by θ, were estimated using the RRs to the 36 items. The estimated attitude scale was normed so that the overall mean was 0 and the variance was 1. For each test item, a and b were estimated. The estimated as were about 1. This indicated that the item parameters were reasonably related to the latent variable and that the items had almost equal weights in estimating the attitudes. Other item properties are discussed below.

The model was extended with a Bayesian ANOVA component to identify correlates of cheating. In this study, three demographic characteristics (gender, year of major, and living arrangements), one academic behavior (time spent studying), and one extracurricular activity (fraternity/sorority membership) were considered as student characteristics that were likely to influence cheating behavior. Table 1 gives the group means for each factor. Conducting a multifactor Bayesian ANOVA was not possible because the data set was too small to investigate simultaneously all main effects and (higher level) interactions. Note that the Bayesian ANOVA model also required constancy of the error variance in each group.

In the proposed parametric Bayesian ANOVA approach, several assumptions were made. It was assumed that attitudes and the transformed response probabilities were independently drawn from normally distributed populations, the populations had the same variance, and the means were linear combinations of factor effects. In this real data example, violations of these assumptions were investigated. Within the Bayesian framework, a residual analysis was performed to investigate nonconstancy of the error variance, nonindependence of the error terms, outliers, and nonnormality of the error terms. Gelman et al. (1995, chap. 8) contains an overview of ways to do diagnostic checks for the linear regression model. Both linear models were investigated that were parts of the entire model: (a) the linear model presented as a normal population distribution for the attitudes with variance σ^2 and mean $\mathbf{x}\boldsymbol{\lambda}$ and (b) the linear model in equation (14). No serious departures were found. Some posterior distributions of residuals showed slightly heavier tails than a normal distribution. This was caused by a relatively large group of students with low response probabilities of positive responses to certain items.

The estimated mean attitude for females was slightly above zero and for males was slightly below zero. This means that the female students cheated slightly more than the male students. However, a single-factor Bayesian ANOVA suggested that the effect of gender was not significant, $P(F(1, 342) \geq 5.93) = .015$. Cheating was uncorrelated with year in college, $P(F(5, 342) \geq 2.52) = .030$; however, freshmen cheated less than others. Interpretation of the relationship between cheating and year of major was difficult because many characteristics (such as motivation, age, and experience) change as students progress through the grade levels. For example, cheating is known to be negatively correlated with age. Finally, attitudes toward cheating differed significantly across majors, $P(F(6, 341) \geq 8.01) < .001$. The largest difference was found between CS and EST students, with EST students more inclined to cheat than CS students.

Analysis at the item level. Besides obtaining information at the group level, IRT modeling provides information at the item level. Table 2 gives the true percentages of positive answers using the item response function (IRF) approach for the 36 items that constitute the attitude scale. The estimates of the true proportion of positive responses for items $k = 1, \ldots, K$ are given in Table 2

Table 1
Group Mean Attitudes Toward Cheating

Characteristic	Group	*n*	Mean Attitude
Demographic	Gender		
	Male	229	−.072
	Female	120	.137
	Year of major		
	First	52	−.309
	Second	73	.093
	Third	66	.131
	Fourth	61	.018
	Fifth	45	.009
	>Fifth	52	.025
	Living arrangements		
	Campus	83	−.207
	City	40	.059
	Parents	226	.086
Academic behavior	Time spent studying (in hours)		
	<10	49	.190
	10-20	99	.112
	20-30	117	.030
	>30	84	−.201
Extracurricular activity	Fraternity/sorority membership		
	Passive	246	−.046
	Active	103	.110
Situational factor	Major		
	CS	50	−.546
	PAT	53	.232
	ACS	53	.337
	ST	46	.065
	EST	66	.115
	ME	49	.136
	PS	32	.025

Note. CS = computer science; PAT = public administration and technology; ACS = applied communication sciences; ST = science and technology; EST = educational science and technology; ME = mechanical engineering; PS = philosophy of science.

(cf. columns 3 and 4). Table 2 shows that per item the mean probability of a positive response under the normal ogive model was close to the maximum likelihood estimate of the true proportion of positive responses via a standard binomial analysis of the RR data using equation (3). Note that in equation (3) a maximum likelihood estimate was obtained for the true proportion of positive responses to an item. On the other hand, the expected true proportion of positive responses equaled the probability of a positive true response under a normal ogive model (e.g., see Baker, 1992, p. 27; Glas, 1989). As a result, similar percentages were as expected in Table 2.

Note that this standard analysis does not allow an individual response model for a single observation. It cannot be used to make inferences concerning individual respondents or item characteristics. The standard method can be used only for estimating the true proportion of positive responses in

Table 2
Items Measuring Cheating Behavior

	Item	% Reporting Behavior	
		Traditional	IRT
	During an exam or a test:		
1	Sent information using signals or sign language	.04	.07
2	Received information using signals or sign language	.04	.05
3	Tried to confer with other students	.20	.20
4	Allowed others to copy your work	.31	.32
5	Others allowed you to copy their work	.23	.24
6	Switched exams to copy answers	.00	.00
7	Obtained information outside the classroom or examination area	.04	.00
8	Used cell phones to exchange information	.00	.00
9	Used a programmed calculator to retrieve information	.09	.04
10	Used crib notes or cheat sheets	.40	.39
11	Used unauthorized material such as books or notes	.24	.23
12	Looked at another student's test paper with his or her knowledge	.24	.24
13	Looked at another student's test paper when walking by	.02	.00
14	Tried to retrieve information from a supervisor	.06	.05
15	Tried to retrieve information from exams handed in	.00	.00
16	Placed books or notes outside the classroom or examination area to cheat	.01	.01
17	Added information to authorized material	.38	.36
18	Made unauthorized use of the Internet	.05	.01
19	Illicitly gained advance information about the contents of the exam	.03	.00
20	Took along an exam illegally	.26	.24
21	Delayed taking an examination to cheat	.00	.00
22	Lied to obtain another opportunity for taking a test	.00	.00

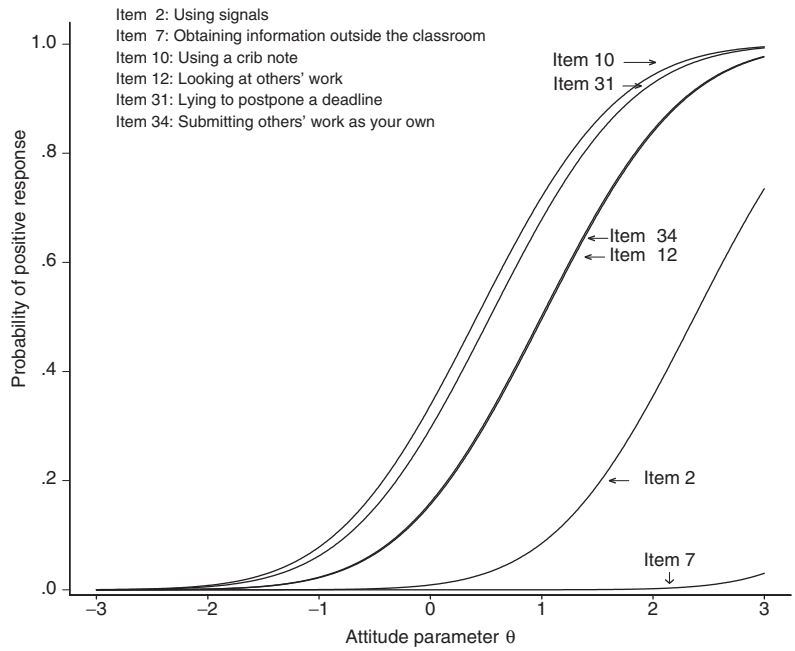
(continued)

Table 2 (continued)

	Item	% Reporting Behavior	
		Traditional	IRT
23	Tried to change results after handing in the exam	.01	.00
24	Lied to hand in the exam at another time	.00	.00
25	Lied about handing in the exam	.02	.00
26	Took an examination for someone else	.00	.00
27	Let someone else take an examination for you	.00	.00
28	Falsified certificates and/or grades	.00	.00
29	Invented data (i.e., entered nonexistent results into the database)	.26	.23
30	Altered data to obtain significant results	.28	.27
31	Lied to postpone a deadline	.37	.36
32	Lied about submitting course work	.01	.01
33	Minimized effort in a joint assignment	.23	.20
34	Submitted course work from others without their knowledge	.25	.24
35	Turned in a paper obtained in large part from a Web site	.17	.15
36	Paraphrased material from another source without acknowledging the author	.16	.13

Note. IRT = item response theory.

Figure 1
Item Characteristic Functions for Ways of Cheating

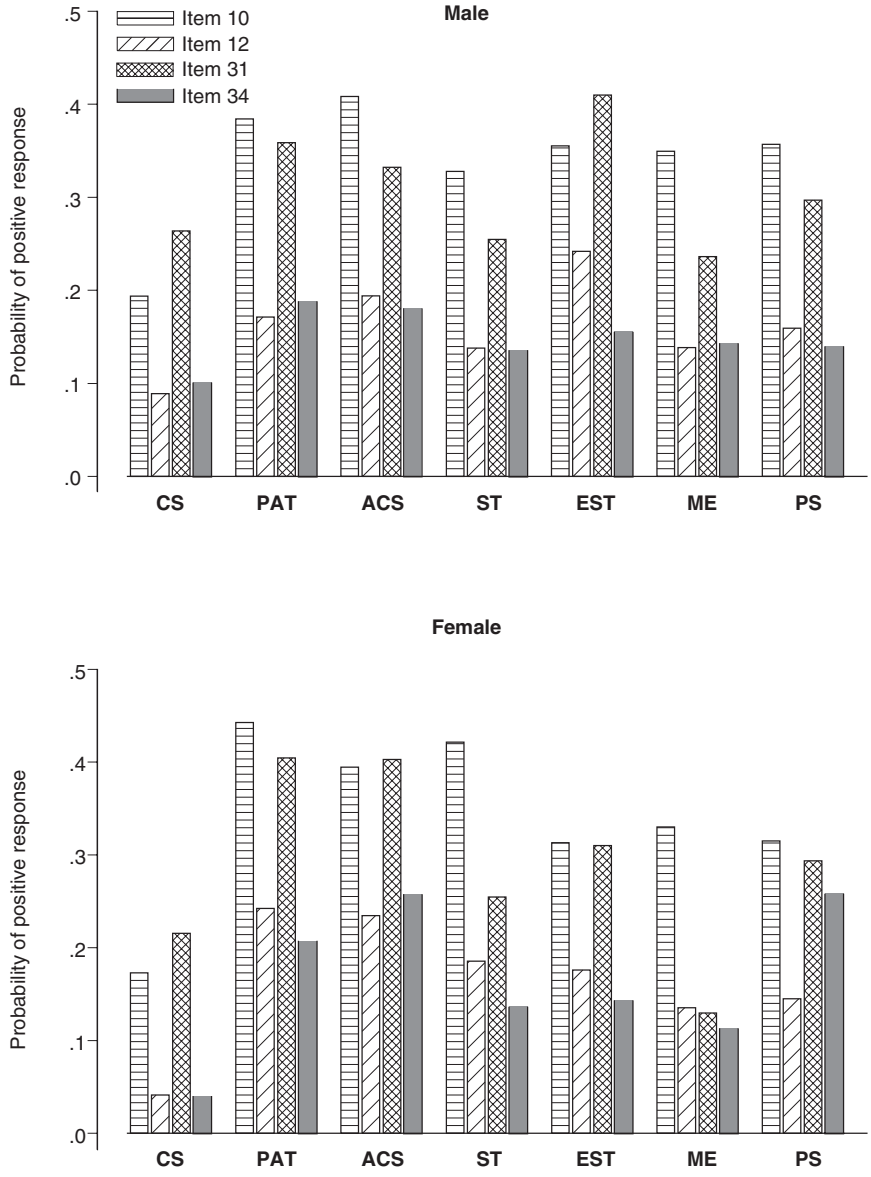


the population. The strength of extending the RR method with an IRT model is the additional information that can be obtained at the individual and item levels.

To illustrate the advantages of the IRT approach, Figure 1 presents the IRFs of 6 items that measured different ways of cheating. These IRFs revealed that the items had different threshold values. The IRFs of Item 10 (“use of crib notes”) and Item 31 (“lying to postpone a deadline”) had the lowest thresholds and thus represented the most popular ways of cheating from the 6 selected items (they were also the most popular methods of all 36 items). In contrast, the IRFs of Item 2 (“using signals”) and Item 7 (“obtaining information outside the classroom”) had higher thresholds and were thus less popular. Note that here the term “popular” refers to a preference of those respondents who are likely to cheat and does not refer to a general preference of the population of respondents.

Most interesting, however, is that the IRFs (Figure 1) supported the conclusion that using signals was a popular method for students with high cheating attitudes (i.e., high θ values). These conclusions cannot be drawn from a traditional RR data analysis. In fact, based on a traditional analysis, one would conclude that the use of signals is not a popular method. In contrast, the IRT analysis showed that students who use signals may be frequent cheaters. Figure 2 displays the group mean probabilities of positive responses to Items 10, 12, 31, and 34 categorized by gender and major. Items 2 and 7 were dropped because all corresponding group mean probabilities were almost zero. The probability of positive response differences per group was investigated. On the basis of visual inspection, gender differences can be neglected for these four items, a conclusion also supported by the Bayesian ANOVA that found the group means did not differ significantly using the parametric and nonparametric approaches.

Figure 2
Group Mean Probabilities of Positive Responses to Items 10, 12, 31, and 34



Note. CS = computer science; EST = educational science and technology; PS = philosophy of science; ME = mechanical engineering; PAT = public administration and technology; ST = science and technology; ACS = applied communication sciences.

Discussion

A new method was proposed that combined the RR model with an IRT model. As a result, information was obtained at the individual level given randomized item responses. In general, the probability of a positive response, regarding the sensitive question, was modeled by an IRT model. In this way,

individual attitudes could be estimated given randomized item responses. That is, attitudes of individuals toward a sensitive topic were estimated while maintaining privacy regarding their individual answers. The incorporation of an IRT model also enabled the possibility for other statistical analyses such as Bayesian ANOVA. The Bayesian ANOVA was integrated within the IRT model for analyzing RR data. Thus, mean group attitude differences of respondents across groups could be tested.

In the forced RR sampling design, it is possible to estimate the model parameters in a frequentist framework via a marginal maximum likelihood method. Some effort has to be made to take account of other RR sampling designs. More research needs to be done to explore maximum likelihood methods that are able to estimate simultaneously all model parameters including the extensions to model group differences in attitudes and response probabilities. One option to be explored is the generalized nonlinear mixed model approach described in De Boeck and Wilson (2004). It should be noted that a maximum likelihood estimation method relies on asymptotic results that might be unrealistic and that are typically not satisfied when the number of individuals is rather low. Rupp, Dey, and Zumbo (2004) concluded that Bayesian IRT estimates are generally closer to the true values and less variable than maximum likelihood estimates when the number of items and/or persons is low. Other advantages of Bayesian parameter estimation are that (a) it handles perfect and imperfect scores, (b) it allows a judgment of possible parameter values, and (c) it easily handles restricted parameter values that are typical for IRT models. The proposed MCMC estimation procedure is very flexible and can be easily adjusted to handle different RR sampling designs, more levels of random effects (i.e., not limited to two-level data), and crossed random effects, among others. The MCMC software is available via the first author's Web site (http://users.edte.utwente.nl/fox) and contains a set of functions that can be used within R.

This new method was applied to collect and analyze data that shed some light on cheating behavior in college. Direct questioning (the confidential questionnaire) may lead to biased results because students are asked to be honest about their own dishonesty. The RR technique was used to obtain more reliable answers on sensitive questions in combination with an IRT model to obtain information at the individual level.

Another important contribution of the combined model was the analyses at the individual and item levels. Using IRT, the threshold values were determined for 36 items that measured ways of cheating. Popular methods were the use of cribs and copying work from others. Depicting IRFs showed that items with high threshold values were endorsed only by students who cheat often. Only 25% of all students, but 72% of the male and 49% of the female students belonging to the group with high attitude values, admitted cheating. Academic dishonesty thus seems to be quite a common practice for a selected group of students. Students who are part of this group need more attention to influence the academic integrity of the academic environment.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, *17*, 251-269.

Anderman, E. M., & Midgley, C. (2004). Changes in self-reported academic cheating across the transition from middle school to high school. *Contemporary Educational Psychology*, *29*, 499-517.

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison-Wesley.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum.

Davis, S. F., Grover, C. A., Becker, A. H., & McGregor, L. N. (1992). Academic dishonesty: Prevalence, determinants, techniques, and punishments. *Teaching of Psychology*, *9*, 16-20.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models*. New York: Springer.

Donovan, J. J, Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance*, *16*, 81-106.

Fowler, F. (2002). *Survey research methods* (3rd ed.). Thousand Oaks, CA: Sage.

Fox, J. A., & Tracy, P. E. (1986). *Randomized response*. Beverly Hills, CA: Sage.

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

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

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall.

Glas, C. A. W. (1989). Extensions of the partial credit model. *Psychometrika*, *54*, 635-659.

Horvitz, D. G., Shah, B. V., & Simmons, W. R. (1967). The unrelated question randomized response model. In *Proceedings of the Social Statistics Section* (pp. 65-72). Washington, DC: American Statistical Association.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. New York: Chapman & Hall/CRC.

Mislevy, R. J., & Bock, R. D. (1990). BILOG 3. Item analysis and test scoring with binary logistic models (2nd ed.) [Computer software and manual]. Mooresville, IN: Scientific Software.

Murdock, T. B., Miller, A., & Kohlhardt, J. (2004). Effects of classroom context variables on high school students' judgments of the acceptability and likelihood of cheating. *Journal of Educational Psychology*, *96*, 765-777.

Newstead, S. E., Franklyn-Stokes, A., & Armstead, P. (1996). Individual differences in student cheating. *Journal of Educational Psychology*, 88, 229-241.

Patz, R. J., & Junker, B. W. (1999a). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*, 342-366.

Patz, R. J., & Junker, B. W. (1999b). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146-178.

Raftery, A. E. (1995). Bayesian model selection in social research. *Social Methodology*, *25*, 111-163.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271-282.

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology*, *44*, 75-92.

Rupp, A. A., Dey, D. K., & Zumbo, B. D. (2004). To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to modeling. *Structural Equation Modeling*, *11*, 424-451.

Scheers, N. J., & Dayton, C. M. (1987). Improved estimation of academic cheating behavior using the randomized response technique. *Research in Higher Education*, *26*, 61-69.

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.

Whitley, B. E. (1998). Factors associated with cheating among college students: A Review. *Research in Higher Education*, *39*, 235-274.

Acknowledgments

The authors thank Esther Cohen and Elizabeth A. Verhoeff for providing the empirical data used in this study.

Author's Address

Address correspondence to Jean-Paul Fox, University of Twente, Department of Research Methodology, Measurement and Data-Analysis, P.O. Box 217, 7500 AE Enschede, Netherlands; e-mail: fox@edte.utwente.nl.