Journal of Educational and Behavioral Statistics
http://jebs.aera.net

A Mixed Effects Randomized Item Response Model

J.-P. Fox and Cheryl Wyrick

JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS published online 26 November 2007
DOI: 10.3102/1076998607306451

Journal of Educational and Behavioral Statistics
Month XXXX, Vol. XX, No. X, pp. xx–xx
DOI: 10.3102/1076998607306451

A Mixed Effects Randomized Item Response Model

J.-P. Fox
University of Twente

Cheryl Wyrick
Tanglewood Research

The randomized response technique ensures that individual item responses, denoted as true item responses, are randomized before observing them and so-called randomized item responses are observed. A relationship is specified between randomized item response data and true item response data. True item response data are modeled with a (non)linear mixed effects and/or item response theory model. Although the individual true item responses are masked through randomizing the responses, the model extension enables the computation of individual true item response probabilities and estimates of individuals' sensitive behavior/attitude and their relationships with background variables taking into account any clustering of respondents. Results are presented from a College Alcohol Problem Scale (CAPS) where students were interviewed via direct questioning or via a randomized response technique. A Markov Chain Monte Carlo algorithm is given for estimating simultaneously all model parameters given hierarchical structured binary or polytomous randomized item response data and background variables.

Keywords: *item response theory model; MCMC; mixed effects; randomized response data*

Introduction

In survey sampling, interest is often focused in obtaining information about highly personal sensitive questions. Direct questioning of such sensitive questions leads to refusals, nonresponses, or socially desirable answers. Different techniques have been developed to improve the quality of such survey data. Methods like emphasizing confidentiality and/or anonymity of response have not been shown to encourage greater cooperation by respondents such that the quality of the survey data improves. However, one method introduced by Warner (1965), the randomized response (RR) technique, can outperform direct ways of asking sensitive questions (Lensvelt, Hox, van der Heijden, & Maas, 2005). Each respondent uses a randomization device to select a question from two complementary questions, and the respondent answers the selected question. The interviewer will not know which question is being selected, so the respondent is protected. Greenberg, Abul-Ela, Simmons, and Horvitz (1969) showed that an unrelated question model could improve Warner's technique. This unrelated question model could be built into the randomization device. For instance, the randomization device could be a die. The respondent would be asked to throw the die and answer no if the outcome is 1, answer yes if the outcome is 6, and answer the sensitive question in the other situations.

Two broad classes of RR sample designs will be considered, related question and unrelated question designs. The related question design applies to a population with two classes. Warner (1965) showed that among others, maximum likelihood estimates as well as confidence intervals can be obtained of the true population proportion of respondents positively answering the sensitive questions. Similar kinds of information can be obtained from the related question design. However, the unrelated question design is easily extended to the multiproportions case (Greenberg et al., 1969). Besides the apparent usefulness of the traditional RR technique (e.g., Greenberg et al., 1969; Warner, 1965), inferences from applications utilizing them are limited to estimating population proportions (e.g., rates of criminal behavior, Tracy & Fox, 1981; rates of academic cheating, Scheers & Dayton, 1988). Often it is of interest to investigate and relate factors underlying the sensitive characteristics. In academic cheating studies (Cizek, 1999), interest is focused on identifying factors (e.g., student characteristics, demographic characteristics, and academic behavior) as possible correlates of cheating. In the present article, results are presented from a College Alcohol Problem Scale (CAPS). One of the objectives is to establish factor-derived scales and background characteristics that are related to alcohol dependence given polytomous randomized responses. In sample surveys, respondents are often nested within groups and responses from members of one group are likely to be correlated. To assess the relationship between the responses and some factors, this dependency must be taken into account. The usual randomized response models do not allow a hierarchical analysis of the RR data. Statistical methods for hierarchical structured data, like analysis of variance or multilevel analysis, cannot be applied because only polytomous randomized responses are observed.

Attempts have been made to develop methods for analyzing RR data given individual and/or group characteristics. The logistic regression modeling approach of Maddala (1983) and Scheers and Dayton (1988) has been extended to model individual binary response probabilities to items and their relationship with individual background variables (e.g., Lensvelt, van der Heijden, & Laudy, 2006). Recently, new methods were developed to analyze multivariate binary RR data where it can be assumed that an underlying latent variable is causally related to the observed indicators. Fox (2005b) developed a class of randomized item response theory (IRT) models for binary data within a Bayesian framework. These models can be used to estimate relationships between a latent variable and background variables at the level of individuals or groups given randomized and/or directly observed binary item responses. Böckenholt and van der Heijden (2007) developed a class of models within a frequentist framework for binary RR data and individual background information. They considered an interesting extension that takes account of response bias caused by respondents who do not follow the RR instructions and/or respondents who do not trust the RR protection mechanism and give a negative response regardless of the question being asked. Fox (in press) proposed a novel Beta-Binomial model for analyzing multivariate individual count data observed via a randomized response sampling design. This model allows for the estimation of individual response probabilities (response rates) taking account of a clustering of respondents using an empirical Bayes approach. The model is particularly useful for small data sets. However, the model assumes a constant individual response rate across items and equal item difficulties. Although the model is proven to be quite robust against these model violations, when a larger data set is available a less restrictive model may fit the data better and may lead to more accurate inferences.

In the present article, several model extensions will be discussed for individuals nested in groups given binary or polytomous univariate or multivariate RR data. The approach is focused on models for the true item response that is not observed because this response is randomized before it is observed. The true item responses are estimated based on the assumption that the respondents follow the RR design. The respondent's true position on the sensitive questions can only be approximated when respondents do not follow the RR instructions. A probabilistic model is defined that relates the observed randomized response with the unobserved true item response. Then, the true item response is modeled with a (non)linear mixed effects and/or item response theory model. This mixed effects randomized item response model improves the statistical analyses of RR data. It enables the measurement of individual true item response probabilities and the measurement of individual (sensitive) latent characteristics. The individual latent characteristics and true item response probabilities and their relationship with explanatory variables can be estimated taking into account any clustering of respondents.

In the following, the related and unrelated randomized response designs will be discussed for binary and polytomous (ordinal) RR data. Then, a mixed effects model will be introduced that relates explanatory variables with individual true item response probabilities. Subsequently, different IRT models will be introduced for modeling the true item responses that are linked via a randomized response model to the observed randomized responses. After a simulation study, an application will be given of the analysis of hierarchical structured RR data from survey items, known as the CAPS instrument, measuring alcohol dependence among college students. In the last part, a Markov Chain Monte Carlo (MCMC) algorithm will be presented for estimating simultaneously all parameters.

Related and Unrelated Randomized Response Designs

Warner (1965) introduced the concept of randomized response. A respondent randomly selects one of two statements of the form: (1) I have the sensitive characteristic and (2) I do not have the sensitive characteristic. A randomization device (e.g., throwing a coin or a die) is used to select one of the two related questions. The respondent answers true or false without revealing which question was selected by the randomizing device. Let p denote the probability that Question 1 will be selected by the randomizing device. Let π denote the population proportion with the sensitive characteristic. Then, the probability that respondent indexed i gives a positive response equals,

$$P(y_i = 1) = \pi p + (1 - \pi)(1 - p). \tag{1}$$

In the unrelated question design (Greenberg et al., 1969), the second question is not related and completely innocuous and the probability of a positive response is known. The unrelated question can also be built into the randomizing device. Then, two probabilities are specified by the randomizing device, probability p_1 that the respondent has to answer the sensitive question and the conditional probability p_2 of a positive response given a forced response (Edgell, Himmelfarb, & Duchan, 1982). The probability of a positive response can be stated as,

$$P(y_i = 1) = \pi p_1 + (1 - p_1)p_2. \tag{2}$$

The extension to more than two response categories is easily made to multiple, say $c = 1, \ldots, C$, response categories. Let $\pi(c)$ denote the proportion of respondents scoring in category c, and the randomization device determines if the item is to be answered honestly with probability p_1 or a forced response is scored with probability $1 - p_1$. If a forced response is given, it is scored in category c with probability $p_2(c)$, $c = 1, \ldots, C$. Then, the probability of observing a score in category c equals

$$P(y_i = c) = \pi(c)p_1 + (1 - p_1)p_2(c). \tag{3}$$

Note it is assumed that the response probability of scoring in category c of the nonsensitive unrelated question is known a priori. This is more efficient because it reduces the sampling variability. Furthermore, it is quite easy to define unrelated neutral questions whose response probabilities are known in advance (e.g., Greenberg et al., 1969). In a setting where the response probabilities of the unrelated question are unknown, two independent distinct samples, say A_1 and A_2, are needed from the population where two independent randomization devices are employed. Let the randomization devices be such that p_1^1 is the probability that the respondent has to answer the sensitive question in A_1 and p_1^2 in A_2. Subsequently, given group membership the probability of scoring in category c equals

$$P(y_i = c) = \begin{cases} \pi(c)p_1^1 + \left(1 - p_1^1\right)p_2(c) & \text{if } i \in A_1 \\ \pi(c)p_1^2 + \left(1 - p_1^2\right)p_2(c) & \text{if } i \in A_2. \end{cases} \tag{4}$$

When selecting p_1^1 close to p_1^2 the point estimates of $\pi(c)$ and $p_2(c)$ may be unstable and greater than unity (Greenberg et al., 1969).

Individual Response Probabilities

In general, the randomized response technique is used to estimate the proportion, π, of respondents belonging to a sensitive class in the population. Theoretical details about the estimation of π can be found in, among others, Greenberg et al. (1969) and Warner (1965). If additional information is available per respondent that can be related to the individual's probability of a yes response or scoring in category c, it becomes interesting to model the true individual response probabilities within a randomized response sample design. Maddala (1983) and Scheers and Dayton (1988) incorporated explanatory variables in a randomized response model, so-called covariate randomized response models. They showed improved parameter estimates, a reduction in sampling error, when using covariates that correlate with the sensitive characteristic. In the same way, when modeling the true individual response probabilities, the sampling error of the corresponding estimates can be reduced. This may improve estimates of the group-specific proportions with the sensitive characteristic in the population depending on the available explanatory information because each group specific proportion estimate is constructed from a weighted average of a group-specific estimate and the sample estimate where the weights are specified by the corresponding standard errors. Moreover, it will provide information at the individual level because estimates of the individual response probabilities and their relationship with the explanatory variables will be obtained.

Assume that there are $j = 1, \ldots, J$ groups and $i = 1, \ldots, n_j$ individuals nested within each group. Let y_{ijk} denote the randomized response for an individual, indexed ij, to an item indexed k. Subsequently, $\widetilde{y}_{ijk}$ denotes the true item response in the randomized response design. Assume that the probability of observing a response in category $c = 1, \ldots, C_k$ is modeled according to Equation 3. Another variable is defined: $H_{ijk} = 1$ when for respondent ij the randomizing device determines that item k is to be answered truthfully and $H_{ijk} = 0$ otherwise. It follows that $P(H_{ijk} = 1) = p_1$ and in that case the randomized response equals the true response, that is, $y_{ijk} = \widetilde{y}_{ijk}$ when $H_{ijk} = 1$. Now, the probability that individual ij scores a true response in category c' given that a randomized response is observed in category c can be derived:

$$P\left(\tilde{y}_{ijk} = c' \mid y_{ijk} = c\right) = \frac{P\left(\tilde{y}_{ijk} = c', y_{ijk} = c\right)}{P\left(y_{ijk} = c\right)} \tag{5}$$

$$= \frac{\sum_{l \in (0,1)} P\left(\tilde{y}_{ijk} = c', y_{ijk} = c \mid H_{ijk} = l\right) P\left(H_{ijk} = l\right)}{\sum_{l \in (0,1)} P\left(y_{ijk} = c \mid H_{ijk} = l\right) P\left(H_{ijk} = l\right)} \tag{6}$$

$$= \frac{\pi_{ijk}(c')p_1 I(c = c') + \pi_{ijk}(c')p_2(c)(1 - p_1)}{\pi_{ijk}(c)p_1 + p_2(c)(1 - p_1)} \tag{7}$$

Here it is assumed that the true response and the randomized response are independent when a randomized response is to be given. This assumption is not valid when respondents do not follow the instructions corresponding to the RR design. Equation 7 can be presented as:

$$P\left(\tilde{y}_{ijk} = c' \mid y_{ijk} = c\right) = \begin{cases} \dfrac{\pi_{ijk}(c')\left(p_1 + (1 - p_1)p_2(c)\right)}{\pi_{ijk}(c)p_1 + (1 - p_1)p_2(c)} & \text{if } c = c' \\[2ex] \dfrac{\pi_{ijk}(c')(1 - p_1)p_2(c)}{\pi_{ijk}(c)p_1 + (1 - p_1)p_2(c)} & \text{if } c \neq c'. \end{cases} \tag{8}$$

In summary, a relationship is established between an observed randomized response and a true (latent) response. It turns out that this functional relationship between the observed and true response data allows the specification of individual response probabilities.

When the probability of scoring in a specific category of the unrelated question is unknown Equation 7 is to be adjusted. The right-hand side of Equation 7 changes slightly by taking into account whether respondent i belongs to group A_1 or A_2 according to Equation 4. A latent variable y^*_{ijk} is defined that presents the latent response to the unrelated question k of respondent ij. Subsequently, the conditional probability $P\left(y^*_{ijk} = c' \mid y_{ijk} = c\right)$ needs to be specified that relates the observed randomized response with the latent response to the unrelated question. This conditional probability is derived in the same way as in analyzing the conditional probability $P\left(\tilde{y}_{ijk} = c' \mid y_{ijk} = c\right)$. Note that latent variable $\tilde{y}_{ijk}$ presents the latent response to related question k. It follows that the latent response vector $\mathbf{y}^*$ is multinomial distributed with cell probabilities $p_2(c)$, $c = 1, \ldots, C_k$. The conditional posterior distribution of the cell probabilities $p_2(c)$ given the latent responses $\mathbf{y}^*$ is Dirichlet when using a conjugated Dirichlet prior distribution. Latent responses to the unrelated question and cell probability values are easily sampled from the conditional posterior distributions within an MCMC algorithm. However, this procedure will not be emulated as this procedure is statistically inefficient and it is not difficult to construct a randomization device such that the response probabilities to the unrelated questions are known.

The Model

Probit and Logistic Response Functions

Through the relation between the true and randomized response data, the vector of true responses $(\widetilde{\mathbf{y}})$ can be modeled. Suppose $\widetilde{y}_{ijk}$ denotes a binary outcome and let z_{ijk} be a continuous latent variable such that $\widetilde{y}_{ijk} = 1$ if z_{ijk} is positive and $\widetilde{y}_{ijk} = 0$ if z_{ijk} is negative. A probit model is defined as

$$\pi_{ijk} = P(\widetilde{y}_{ijk} = 1) = \Phi(z_{ijk}), \tag{9}$$

where $\Phi(.)$ represents the cumulative normal distribution function. A logistic response function can be assumed, then the logistic function $L(.)$ replaces $\Phi(.)$.

Here, attention is focused on polytomous ordinal data, but other polytomous responses can be handled in a similar way by using the proper response model for analyzing the data. Assume that $\widetilde{y}_{ijk}$ denotes a categorical outcome and z_{ijk} the underlying latent score such that the probability of individual ij scoring in category $c = 1, \ldots, C$ equals

$$\pi_{ijk}(c) = P\left(\widetilde{y}_{ijk} = c \mid \boldsymbol{\kappa}, z_{ijk}\right) = \Phi\left(z_{ijk} - \kappa_{k,(c-1)}\right) - \Phi\left(z_{ijk} - \kappa_{k,c}\right) \tag{10}$$

or replace $\Phi(.)$ with

$$L(z_{ijk} - \kappa_{k,c}) = \frac{1}{1 + \exp[-(z_{ijk} - \kappa_{k,c})]} \tag{11}$$

where $\boldsymbol{\kappa}$ are the threshold parameters such that $\kappa_{k,r} > \kappa_{k,s}$ whenever $r > s$, with $\kappa_{k,0} = -\infty$ and $\kappa_{k,C} = \infty$.

Multiple Item Responses

The respondents answer to a series of multiple items indexed $k = 1, \ldots, K$. It will be assumed that the items are composed to measure some underlying attitude. The true categorical outcome, $\widetilde{y}_{ijk}$, represents the item response of person ij on item $k(k = 1, \ldots, K)$. These item responses may be dichotomous or polytomous. Let θ_i denote the latent abilities or attitudes of the respondents responding to the K items. They are collected in the latent vector $\boldsymbol{\theta}$. For dichotomous item responses a two-parameter IRT model is used for specifying the relation between the level on a latent variable and the probability of a particular item response. That is

$$P\left(\widetilde{y}_{ijk} = 1 \mid \theta_{ij}, a_k, b_k\right) = \Phi\left(a_k\theta_{ij} - b_k\right), \tag{12}$$

where a_k is the item discrimination parameter, and b_k is the item difficulty parameter. The item parameters will also be denoted by $\boldsymbol{\xi}_k$, with $\boldsymbol{\xi}_k = (a_k, b_k)$. For polytomous item responses, the probability that an individual obtains a grade $c(c = 1, \ldots, C_k)$ on item k is defined by a graded response model (GRM) described by Samejima (1969)

$$P\left(\tilde{y}_{ijk} = c \mid \theta_{ij}, a_k, \boldsymbol{\kappa}_k\right) = \Phi\left(a_k\theta_{ij} - \kappa_{k,(c-1)}\right) - \Phi\left(a_k\theta_{ij} - \kappa_{k,c}\right) \tag{13}$$

where the boundaries between the response categories are represented by an ordered vector of thresholds κ. In this case, let $\boldsymbol{\xi}_k = (a_k, \kappa_k)$. Consequently, there are a total of $C_k - 1$ threshold parameters and one discrimination parameter for each item. For the logistic IRT model replace $\Phi(.)$ with $L(.)$.

Linear Mixed Effects Model

Individual response probabilities regarding the true outcomes given randomized response data can be modeled as a function of some explanatory variables. Furthermore, effects of group-level variables on the individual's binary or ordinal true response may vary across groups. The latent continuous response can be modeled as function of incidence matrices $\mathbf{x}_{ij}$ and $\mathbf{w}_{ij}$ of order $1 \times q$ and $1 \times p$, respectively, as follows:

$$z_{ijk} = \mathbf{w}_{ij}^t \gamma + \mathbf{x}_{ij}^t \boldsymbol{\beta}_j + e_{ijk} \tag{14}$$

where $\mathbf{x}_{ij}$ is the design vector for the random effects. The q-dimensional vector $\boldsymbol{\beta}_j$ contains the random effects, and their distribution is assumed to be multivariate normal with mean zero and covariance matrix T. Furthermore, $\boldsymbol{\gamma}$ is a p-dimensional vector of fixed effects, and the residuals e_{ijk} have mutually independent normal distributions with mean zero and variance σ_e^2. There is independence between random effects of different groups, and the random effects are independent of the residuals e_{ijk}. The covariance matrix $\boldsymbol{T}$ is of dimension $q \times q$.

The mixed effects model Equation 14 can be presented as a multilevel model. For example, the relationship between the outcome variable θ (e.g., attitude or ability), group characteristics, and respondents' characteristics can be described via a multilevel model. Therefore, define θ as the outcome in Equation 14. Then this model is partitioned into a Level 1 model

$$\theta_{ij} = \mathbf{w}_{ij}^{t(1)} \boldsymbol{\gamma}^{(1)} + \mathbf{x}_{ij}^{t(1)} \boldsymbol{\beta}_j + e_{ij}, \tag{15}$$

and a Level 2 model,

$$\boldsymbol{\beta}_j = \mathbf{w}_j^{t(2)} \boldsymbol{\gamma}^{(2)} + \mathbf{u}_j, \tag{16}$$

where $\mathbf{w}_{ij}^{(1)}$ and $\mathbf{w}_j^{(2)}$ are the fixed Level 1 and fixed Level 2 covariates, respectively. The Level 1 residuals e_{ijk} have mutually independent normal distributions with mean zero and variance σ_e^2. The random effects $\boldsymbol{\beta}_j$ are influenced by the Level 2 effects and a random component $\mathbf{u}_j$ distributed normally with mean zero and covariance matrix T. Examples of two-stage random effects models that are based on individual and population characteristics can be found in Goldstein (2003) and Snijders and Bosker (1999). It can be seen that Level 1 covariates are included to explain the variation at the individual level, and Level 2 covariates are included to explain variation at the level of groups. So, a two-stage model allows explicit modeling and analysis of between- and within-individual variation. Note that the structural mixed effects model can be extended in several ways. For example, mixed effects models with several nested sources of heterogeneity (e.g., Goldstein, 2003; Longford, 1987) or models with crossed classified random effects (Skrondal & Rabe-Hesketh, 2004) can be considered.

The combination of a randomized response model, an individual response model for true (latent) response data, and a mixed effects model results in a mixed effects randomized item response model. For example, in Figure 1 a path diagram is given of a randomized item response model. The data $\mathbf{y}$ are observed via a randomized response sampling design. The randomized response model parameters $\mathbf{p}$ are known, and via Bayes's theorem a relation can be specified with the true latent response data $\widetilde{\mathbf{y}}$. Within an IRT model, an individual attitude or ability, θ_{ij}, is related to the latent response data. Then, a multilevel model is specified to model the effect of Level 1, $\mathbf{x}_{ij}$, variables and to model group specific effects of Level 2, $\mathbf{w}_j$ variables on the latent variable θ_{ij}.

The mixed effects randomized response model for a single randomized response item, that is, combining the model in Equation 14 with a randomized response model, can be considered as an extension of the generalized linear mixed model (e.g., Hedeker & Gibbons, 1994; Laird & Ware, 1982; Zeger & Karim, 1991). The generalized linear mixed model is useful for analyzing clustered binary or ordinal response data but cannot handle data obtained via a randomized response sampling design.

The mixed effects ordinal model is usually identified by fixing one threshold parameter. The multilevel randomized item response model is identified by fixing the scale of the latent variable θ. This can be done by fixing a threshold and a discrimination parameter or by fixing the mean and variance of the latent variable (e.g., the mean and standard deviation of θ). Both ways of identifying the model lead to the same results but on a different scale depending on the type of restrictions.

Bayesian Inference

MCMC estimation (see e.g., Gelfand & Smith, 1990; Geman & Geman, 1984) is a powerful tool for estimation in complex models. An MCMC procedure can be applied to estimate simultaneously all model parameters. Within the Bayesian analysis, proper uninformative priors are used to facilitate the computation of a Bayes factor (Kass & Raftery, 1995). Simulated values from the

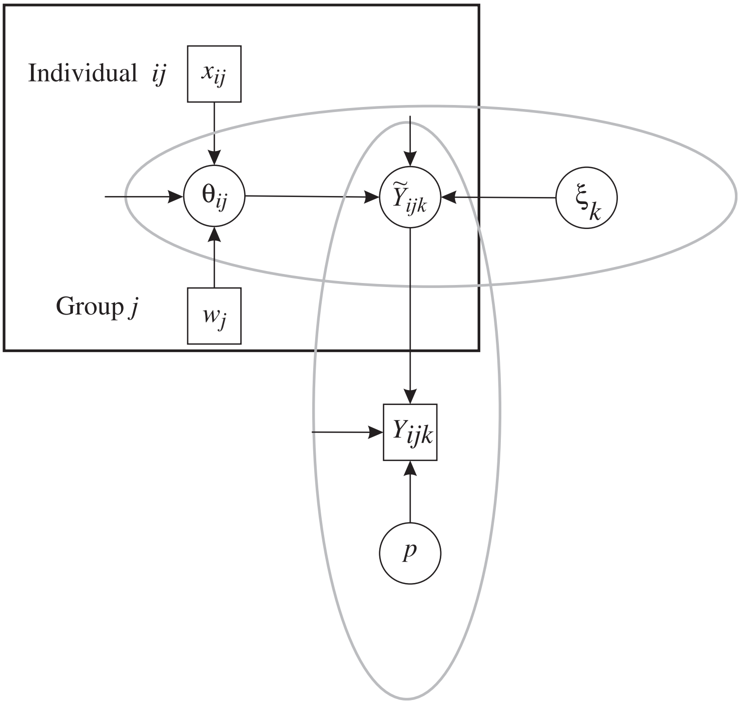

FIGURE 1. *Path diagram of a multilevel randomized item response model.*

posterior distributions are obtained using the Gibbs sampler. The sampled parameter values can be used to estimate all model parameters.

The conditional density of all observations, say multiple item responses, given the parameters, without any hyperprior parameters, equals

$$p(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma_e^2, T) \propto \int p(\mathbf{y} \mid \widetilde{\mathbf{y}}, \boldsymbol{\theta}, \boldsymbol{\xi}) p(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta}, \boldsymbol{\xi}) p(\boldsymbol{\theta} \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma_e^2) p(\boldsymbol{\beta} \mid T) d\widetilde{\mathbf{y}}. \quad (17)$$

The mixed effects randomized item response model contains three components, a randomized response model, $p(\mathbf{y} \mid \widetilde{\mathbf{y}})$, 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, $p(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta}, \boldsymbol{\xi})$, for measuring the underlying attitudes, and a linear mixed effects model, $p(\boldsymbol{\theta} \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma_e^2) p(\boldsymbol{\beta} \mid T)$.

To complete the specification of the model in a Bayesian context proper prior distributions are specified. The joint prior distribution of $(\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma_e^2, T)$ has density:

$$p(\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma_e^2, T) = p(\boldsymbol{\xi})p(\boldsymbol{\theta} \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma_e^2)p(\boldsymbol{\beta} \mid T)p(\boldsymbol{\gamma})p(\sigma_e^2)p(T). \tag{18}$$

For the item parameters, noninformative proper priors for the discrimination and difficulty parameters in the normal ogive model, Equation 12, are used, although other (conjugated) priors can be used as well. So, for $k = 1, \ldots, K$,

$$p(\boldsymbol{\xi}_k \mid \mu_\xi, \Sigma_\xi) = p(a_k, b_k \mid \mu_\xi, \Sigma_\xi) \propto f_{\mathcal{N}(\mu_\xi, \Sigma_\xi)}(a_k, b_k) I(a_k, b_k \in \mathcal{A}), \tag{19}$$

where $\mathcal{A}$ is a sufficiently large bounded interval in $\mathbb{R}$. The hyperprior parameters are specified as $\mu_\xi = (1, 0)$ and $\Sigma_\xi \sim Inv - Wishart(2, \Lambda)$ where $\Lambda = 100 I_2$.

The prior for item parameters in the graded response model, Equation (13), are specified as:

$$p(\boldsymbol{\xi}) = p(\mathbf{a})p(\boldsymbol{\kappa}) \propto \prod_k I(a_k > 0) I(a_k, \kappa_{k,1}, \ldots, \kappa_{k,C_k} \in \mathcal{A}), \tag{20}$$

subject to the condition $\kappa_{k,0} < \kappa_{k,1} < \ldots < \kappa_{k,C_k}$ with $\kappa_{k,0} = -\infty$ and $\kappa_{k,C_k} = \infty$.

According to the specification of the linear mixed effects model, the latent variable $\boldsymbol{\theta}$ is assumed to have a normal distribution with mean $\mathbf{w}\boldsymbol{\gamma} + \mathbf{x}\beta$ and variance σ_e^2. The fixed effects, $\boldsymbol{\gamma}$, are assumed to have independent normal prior, with mean zero and variance σ_γ, and the hyperparameter σ_γ equals a large number that reflects a noninformative prior.

The random effects, $\boldsymbol{\beta}$, have a multivariate normal distribution with mean zero and covariance matrix T, so the conditional prior is,

$$p(\boldsymbol{\beta} \mid T) \propto \exp\Big(-\frac{1}{2}\sum_j \boldsymbol{\beta}_j^t T^{-1} \boldsymbol{\beta}_j\Big). \tag{21}$$

The prior for the covariance matrix T is taken to be an inverse-Wishart density:

$$p(T \mid n_q, S) \propto |T|^{(n_q + q + 1)/2} \exp\Big(-\frac{1}{2} tr\big(ST^{-1}\big)\Big), \tag{22}$$

with unity matrix S and hyperparameter $n_q \geq q$ equal to a small number to specify a diffuse proper prior. The conventional prior for σ_e is the inverted gamma with prior parameters n_1 and s_1 with density

$$p(\sigma_e) \propto \sigma_e^{-\frac{1}{2}(n_1 + 1)} \exp(-\frac{n_1 s_1}{2\sigma_e}). \tag{23}$$

A proper noninformative prior is specified with $s_1 = 1$ and a small value for n_1.

The joint posterior distribution, combining the likelihood in Equation 17 with the specified prior distributions, is intractable analytically, but MCMC methods such as the Gibbs sampler and the Metropolis-Hastings algorithm can be used to draw samples. Then, features of the marginal distributions of interest can be inferred. In Appendix A, details about the full conditional posterior distributions are given in case of the related or unrelated randomized response sampling design and in case of binary or polytomous (ordinal) response data.

Simulation Study

In this section, results are reported from a simulation study based on mixed effects models for randomized response data. In a first example, single randomized item response data were simulated, and in a second example, hierarchical structured multiple randomized item response data (with background variables) were simulated.

Comparison Between Randomized Response Models

A total of, $i = 1, \ldots, N$ binary observations divided at random across $j = 1, \ldots, 10$ groups were generated according to the following mixed effects models,

$$P(\widetilde{y}_{ij} = 1) = \Phi(\gamma_0 + u_{0j} + x_i \gamma_1), \tag{24}$$

where $u_{0j} \sim \mathcal{N}(0, \tau^2)$ and $\mathbf{x}$ values were simulated from a normal distribution with mean zero and standard deviation $1/2$. The true parameter values are given in Table 1. Randomized response data were generated according to Warner's model (related response design) with randomizing proportion $p_1 = 4/5$ and according to the forced response model (unrelated response design) with $p_1 = 4/5, p_2 = 2/3$ given the generated true response data $\widetilde{\mathbf{y}}$.

The MCMC method was used to estimate simultaneously all parameters of the mixed effects model and the mixed effects model combined with Warner's or with the forced response model. The MCMC algorithm was run for 20,000 iterations, convergence was obtained after 5,000 iterations, and the cumulative averages of sampled parameter values resembled the true parameter values.

Table 1 presents the estimates and standard errors for the mixed effects model given the latent response vector $\widetilde{\mathbf{y}}$, labeled under *True Response*, for the mixed effects Warner model, labeled under *Warner*, and for the mixed effects forced response model, labeled under *Forced*. It is apparent that the point estimates resemble the true values for a sample size of $N = 5{,}000$ and the estimates are close to the true values for a sample size of $N = 1{,}000$. The Warner model has the largest estimated standard deviations with respect to parameter γ_1, which was also found by Scheers and Dayton (1988).

TABLE 1
Parameter Estimates of a Mixed Effects Model Given True Response or Randomized Response Data

			True Response		Warner		Forced	
N	Parameter	True	*M*	*SD*	*M*	*SD*	*M*	*SD*
1,000	γ_0	0.00	−0.06	.16	−0.06	.16	−0.07	.16
	γ_1	2.00	1.94	.17	1.81	.26	1.80	.20
	τ	0.25	0.29	.18	0.20	.16	0.24	.16
5,000	γ_0	0.00	−0.12	.15	−0.11	.16	−0.12	.16
	γ_1	2.00	2.01	.07	2.01	.12	1.97	.09
	τ	0.25	0.23	.14	0.22	.15	0.24	.15

Finally, the proportion positive responses π was estimated using M sampled values of latent response data $\widetilde{\mathbf{y}}^{(m)}, m = 1, \ldots, M$ from the MCMC algorithm with

$$\hat{\pi} = \frac{1}{MN} \sum_m \sum_{i,j} \widetilde{y}_{ij}^{(m)}. \tag{25}$$

For $N = 5{,}000$, the simulated proportion of positive responses equals .464. The estimated proportion under the mixed effects Warner model equals .464 with standard deviation .007 and equals .465 with standard deviation .004 under the mixed effects forced response model. These point estimates resemble the estimated proportion using Warner's model (Warner, 1965), $\hat{\pi} = .462$ with standard deviation .010, and using the forced response model (Greenberg et al., 1969), $\hat{\pi} = .464$ with standard deviation .008.

For $N = 1{,}000$, the simulated proportion equals .485 and the estimated proportion under the mixed effects Warner model and the mixed effects forced response model equals .488 (.014) and .480 (.009), respectively, where the standard deviations are given in parentheses. The point estimate under Warner's model and the forced response model equals .490 (.023) and .481 (.019), respectively. For both sample sizes, it can be concluded that the estimated proportions are comparable but that there is a reduction in sampling error due to the specification of a mixed effects model.

Influence of Randomized Responses in a Mixed Effects Analysis

In the present simulation study, parameter estimates of a mixed effects model are compared given directly observed and randomized response data for different randomizing proportions. Suppose that the probability distribution of a (latent) behavior parameter, θ_{ij}, has the same form for each individual ($i = 1, \ldots, n_j$) but the parameters of that distribution vary over $J = 20$ groups $(j = 1, \ldots, J)$. That is, a Level 1 model describes a linear relationship between the behaviors of $N = 1{,}000$ respondents and explanatory variable $\mathbf{x}$ and $\mathbf{w}$, and let a Level 2 model represent the distribution of the random effect parameters, that is,

$$\theta_{ij} = \mathbf{x}_{ij}^{t}\boldsymbol{\beta}_j + w_{ij}\gamma^{(1)} + e_{ij} \tag{26}$$

$$\beta_{0j} = \gamma_0^{(2)} + u_{0j} \tag{27}$$

$$\beta_{1j} = \gamma_1^{(2)} + u_{1j} \tag{28}$$

where $e_{ij} \sim N(0, \sigma^2)$, $\mathbf{u}_j \sim N(0, T)$, and the random effects are independent of the residuals. The first column of $\mathbf{x}$ consists of ones, the second column, and the vector $\mathbf{w}$, contains values generated from a normal distribution with mean zero and standard deviation .30. Polytomous item responses to $K = 10$ questionnaire items, each with three ordinal response categories, were simulated using a graded response model. Polytomous randomized item responses were generated via the unrelated question design with $p_2 = 1/3$. The probability that a truthful response was demanded, p_1, (e.g., "Answer truthfully" was selected by the randomizing device) was considered to be *1* (which resembles a direct response), .80, and .60. A total of 100 data sets were analyzed for each value of p_1. Item discrimination parameter values were sampled as follows, $a_k \sim \log N(\exp(1), 1/4)$. Threshold parameters, κ_{k1}, and κ_{k2}, were sampled from a normal distribution with mean $-1/2$, and $1/2$ (taking order restrictions into account), respectively, and variance $1/4$, for $k = 1, \ldots, K$.

For each data set, the graded response model parameters, and the mixed effects model parameters were estimated simultaneously using 50,000 draws from the joint posterior distribution. The burn-in period consisted of 5,000 iterations. In this simulation study, attention was focused on the mixed effects model parameters. Table 2 presents, for each model parameter, the true set-up value, and the average of the means and standard deviations across the 100 MCMC samples. For identification of the model, the mean and variance of the latent outcome variable were scaled to the true simulated mean and variance, respectively.

It can be seen that there is a close agreement between the true and the average estimated means. For each model parameter, the average of the posterior standard deviations resembled the standard deviation within the 100 estimated posterior means. Note that even for $p_1 = .60$, which means that 40% of the responses were forced responses, the estimated values resemble the true simulated values. The standard deviations of the mixed effect parameter estimates were not increasing due to the incorporation of a randomized response sampling design because for each data set the outcome variables were equally scaled. Furthermore, additional variance in the item parameter estimates due to the randomized response sampling design did not result in biased estimates of the behavior parameters.

TABLE 2
Generating Values, Means, and Standard Errors of Recovered Values

		Direct Response $p_1 = 1$		Forced $p_1 = .80$		Forced $p_1 = .60$	
Parameter	True	*M*	*SD*	*M*	*SD*	*M*	*SD*
Fixed effect							
$\gamma^{(1)}$	0.5	0.50	.03	0.50	.04	.49	.04
$\gamma_0^{(2)}$	0	0.03	.11	0.02	.11	–.08	.10
$\gamma_1^{(2)}$	0	0.04	.11	0.04	.12	.08	.10
Random Effect							
σ_e^2	1.00	1.00	.05	1.00	.05	.98	.05
τ_{00}	0.3	0.36	.08	0.36	.09	.31	.08
τ_{11}	0.2	0.20	.05	0.20	.06	.20	.06
τ_{01}	0	0.13	.04	0.14	.05	.00	.04

College Alcohol Problem Scale

The College Alcohol Problem Scale (CAPS; O'Hare, 1997) was developed to serve as an initial screening instrument for students cited with a first offense for violating their university's rules concerning underage drinking. The items comprising the CAPS scale covered socioemotional problems (hangovers, memory loss, nervousness, depression) and community problems (drove under the influence, engaged in activities related to illegal drugs, problems with the law). Due to the high prevalence of alcohol abuse among college students, it is important that practitioners in student health services or counseling be able to identify students with drinking problems. The randomized response technique was used because of the sensitive character of the survey. As researchers in this area are fully aware, most frequently encountered problems are refusals to respond and intentionally misleading responses designed to conceal undesirable behavior. In this study, it was investigated whether the RR technique improved the accuracy of self-reports of sensitive information, and the beneficial role of incorporating an IRT model for polytomous data in examining the RR data was explored.

In all, 793 student participants from four local colleges and universities, Elon University ($N = 495$), Guilford Technical Community College ($N = 66$), University of North Carolina ($N = 166$), and Wake Forest University ($N = 66$), voluntarily responded to a questionnaire of 16 items in 2002. The questionnaire comprised 3 items that asked participants about their age, gender, and ethnicity (demographic information), followed by the 13 questions of the CAPS instrument, with response categories on a 5-point scale (1 = *never/almost never*, 5 = *almost always*). The CAPS questionnaire is given in Appendix B. A unidimensional latent variable representing alcohol dependence, denoted as θ, was measured by the items, where a higher level indicated that a participant was more likely to have a drinking problem. Each class of participants (5 to 10 participants) was randomly assigned to either the direct questioning (DQ) or the randomized response technique condition. Random assignment at the individual level was not logistically feasible. The 351 students assigned to the DQ condition, denoted as the DQ-group, were instructed to answer the questionnaire as they normally would. They served as the study's control group. The 442 students in the RR condition, denoted as the RR-group, received a spinner to assist them in completing the questionnaire. For each CAPS item, the participant spun the spinner, and wherever the arrow landed determined whether the item was to be answered honestly or dictated the answer choice to be recorded by the participant. The spinner was developed such that 60% of the area was comprised of answer honestly space, and 40% of the area was divided into equal sections to represent the five possible answer choices. Each answer choice was given 8% of the area of the circle, 4% in two different places on the circle. This design resembles the forced response sampling design, Equation 3, with $p_1 = .60$ and $p_2(c) = .20$, for $c = 1, \ldots, 5$. The respondents from the DQ-group and the RR-group were assumed to be selected from the same population.

All response data, obtained via direct questioning and via the randomized response technique, were used to measure the latent behaviors (alcohol dependence) of the respondents on a common scale using the graded response model, Equation 13, combined with the forced randomized response model, where $p_1 = 1$ for DQ responses. This results in the following IRT measurement model for RR data:

$$P(y_{ijk} = c \mid \theta_{ij}, a_k, \boldsymbol{\kappa}_k) = p_1 \left[\Phi\left(a_k \theta_{ij} - \kappa_{k,(c-1)}\right) - \Phi\left(a_k \theta_{ij} - \kappa_{k,c}\right) \right] + (1 - p_1) p_2(c), \quad (29)$$

for $c = 1, \ldots, 5$, $k = 1, \ldots, 13$, and respondents, indexed i, nested in $J = 4$ colleges/universities. It was assumed that the item response functions were the same across groups, that is, the response probabilities at the same alcohol dependence, θ, level did not depend on group membership (in this case, the DQ-group and the RR-group). The model was identified by fixing the mean and variance of the scale of the latent variable to zero and one, respectively. The MCMC algorithm was used to estimate simultaneously all item and behavior parameters using 50,000 iterations with a burn-in period of 5,000 iterations.

Figure 2 shows the item parameter estimates. Several statements can be made concerning the performance of the scale. It can be seen that each CAPS item was able to discriminate. All items had consistently increasing ICC, indicating that the likelihood of endorsing higher levels on each item increased with higher levels of alcohol dependence. The threshold estimates of some items (e.g., Items 9 [nausea or vomiting], 11 [spent too much money on alcohol], and 12 [feeling tired or hung over]) showed that the third response option *sometimes* was more likely to be endorsed in comparison with the other items and can be considered as the less severe items. Items with "high" estimated thresholds values (e.g., Items 7 [hurt

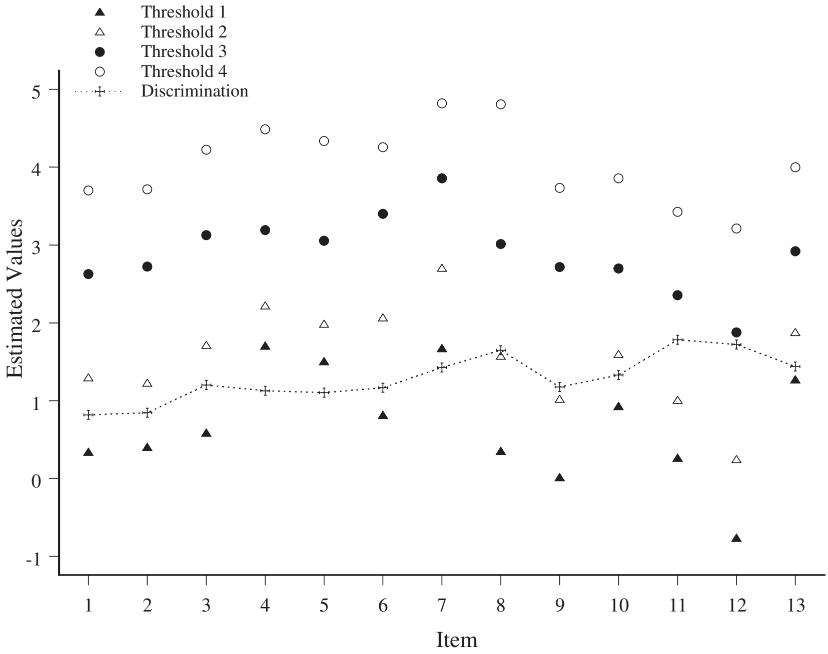

FIGURE 2. *Item parameter estimates given the College Alcohol Problem Scale data.*

another person physically] and 13 [illegal activities associated with drug use]) were severe and they were unlikely to provide significant discrimination in a sample not selected for alcohol dependence. For all items, except Items 11 and 12, the fifth response options (*always*) were not used enough to contribute to the items' discrimination and could in fact be collapsed with the previous option.

Modeling Individual Variation in Alcohol Dependence

A mixture model was defined to model the variation in the respondents' alcohol dependence taking the heterogeneity in the college population into account and to test the hypothesis that in general, students in the RR-group score significantly higher than students in the DQ-group. For both groups, gender ($female = 1$, $male = 0$) and racial origin (Asian, White, Black, other) was administered. In Table 3 gender and ethnicity demographics are given. The overall percentages of gender and ethnicity were similar between the DQ-group and the RR-group.

In an ANOVA approach, the clustering of respondents in a RR-group and DQ-group, the clustering with respect to gender, and the clustering with respect to ethnicity were represented with dummy variables, which resulted in design matrix $\mathbf{w}$. The regression of θ on the known design matrix $\mathbf{w}$ was assumed to be exchangeable across colleges/universities. As a result,

TABLE 3
Gender and Ethnicity Demographics

		Total Count	DQ-Group %	RR-Group %
Gender	Female	274	62	65
	Male	291	38	35
Ethnicity	Asian	22	3	2
	White	650	81	83
	Black	93	12	11
	Other	28	4	12

Note: DQ-Group = direct questioning condition; RR-Group = randomized response condition.

$$\theta_{ij} = \beta_{0j} + \mathbf{w}_j \boldsymbol{\gamma}^{(1)} + e_{ij} \tag{30}$$

$$\beta_{0j} = \gamma^{(2)} + u_j \tag{31}$$

where $e_{ij} \sim N(0, \sigma_e^2)$, and $u_j \sim N(0, \tau^2)$, for $i = 1, \ldots, n_j$, and $j = 1, \ldots, J$. The mixed effects randomized item response model is represented in Equations 29, 30, and 31.

First, the effect of the grouping of respondents in a DQ-group and a RR-group was investigated taking the observed other individual and group differences into account. Indicator variable IRR equals one for students in the RR-group and equals zero otherwise. The DQ-group is the control group and receives no treatment because they responded via direct questioning, whereas the RR-group is the experimental group and receives the treatment because they responded via the randomized response technique. The students were randomly selected in a RR-group and a DQ-group. The estimated value of the parameter corresponding to the indicator variable IRR is .232 and significantly different from zero while controlling for other population differences. This estimate indicates that the RR-group scored significantly higher in comparison to the DQ-group on the standardized alcohol dependence scale. It is to be concluded that the RR technique led to an improved willingness of students in answering truthfully.

The structural relationships between student's alcohol dependence and observed background variables were estimated using the observations from the RR-group because those students were more likely to give honest answers. Besides a mixed effects model, an alternative fixed model was estimated where interest was focused on alcohol dependence of students of the four selected colleges/universities that took part in the experiment and there was no interest in the underlying population. In a similar way, the clustering of students in colleges/universities was represented with the use of dummy variables. The corresponding structural model contained only fixed effects.

In Table 4 the parameter estimates are given for both models. The estimates of the mean and posterior standard deviation of the random effects (universities) are given under the label *mixed effects model*. As was to be expected, these effects are slightly stronger for the fixed model where no random effects are assumed. The estimated variance of the random effects indicates that alcohol dependence of students varies across colleges/universities. However, the corresponding estimated posterior standard deviation is too large for making substantial inferences. The number of clusters is so small that the variation over colleges/universities can also be considered as a fixed effect. Compared with the fixed effects the posterior mean estimates are varying in the same similar way.

It follows that male students scored significantly higher in comparison to female students. That is, male students are more likely to experience alcohol-related problems. There are inequalities in reporting alcohol-related problems across ethnic groups and it turns out that the mean score of Black students is much lower than that of other ethnic groups. With regard to the college/university grouping, the mean score of students from Guilford Technical Community College was higher than the other colleges/universities means. The results indicate that gender, ethnicity, and type of university were associated with alcohol-related problems.

The differences in scores across ethnic groups and colleges/universities were tested using an F test. Let a vector $\boldsymbol{\gamma}_s^{(1)}$ of dimension $J-1$ denote a subset of the p-dimensional vector $\boldsymbol{\gamma}^{(1)}$ with covariate matrix $\mathbf{w}_s$. From Box and Tiao (1973, pp. 125–126) it follows that

$$Q(\boldsymbol{\gamma}_s^{(1)}) = \frac{(\boldsymbol{\gamma}_s^{(1)} - \hat{\boldsymbol{\gamma}}_s^{(1)})\mathbf{w}_s^t\mathbf{w}_s(\boldsymbol{\gamma}_s^{(1)} - \hat{\boldsymbol{\gamma}}_s^{(1)})}{(J-1)s^2} \tag{32}$$

where $s^2 = (\boldsymbol{\theta} - \mathbf{w}\,\hat{\boldsymbol{\gamma}}^{(1)})^t(\boldsymbol{\theta} - \mathbf{w}\,\hat{\boldsymbol{\gamma}}^{(1)})/(N-p)$ is distributed a posteriori as F with $J-1$ and $N-J$ degrees of freedom. The posterior probability can be computed that the point $\boldsymbol{\gamma}_0^{(1)} = 0$ is included in the $(1-\alpha)$ HPD region:

$$\begin{aligned}
p_0 &= P\Big(p(\boldsymbol{\gamma}_s^{(1)} \mid \mathbf{y}) > p(\boldsymbol{\gamma}_0^{(1)} \mid \mathbf{y}) \mid \mathbf{y}\Big)\\
&= \int\!\!\int P\Big(F(J-1, N-J) > Q(\boldsymbol{\gamma}_0^{(1)}) \mid \boldsymbol{\theta}, \boldsymbol{\gamma}_s^{(1)}, \mathbf{y}\Big) p\Big(\boldsymbol{\theta}, \boldsymbol{\gamma}_s^{(1)} \mid \mathbf{y}\Big)\, d\boldsymbol{\gamma}_s^{(1)} d\boldsymbol{\theta}\\
&= \int\!\!\int p_0\Big(\boldsymbol{\gamma}_s^{(1)}, \boldsymbol{\theta}\Big) p\Big(\boldsymbol{\theta}, \boldsymbol{\gamma}_s^{(1)} \mid \mathbf{y}\Big)\, d\boldsymbol{\gamma}_s^{(1)} d\boldsymbol{\theta}\\
&\approx \sum_{\boldsymbol{\gamma}_{(m)}^{(1)}, \boldsymbol{\theta}_{(m)}} p_0\Big(\boldsymbol{\gamma}_{(m)}^{(1)}, \boldsymbol{\theta}_{(m)}\Big)/M,
\end{aligned} \tag{33}$$

where $\left(\theta_{(m)}, \boldsymbol{\gamma}_{(m)}^{(1)}\right)$, $m = 1, \ldots, M$, are MCMC samples from their marginal posterior distribution. Both null hypotheses, $\gamma_3^{(1)} = \gamma_4^{(1)} = \gamma_5^{(1)} = \gamma_6^{(1)}$ and $\gamma_7^{(1)} = \gamma_8^{(1)} = \gamma_9^{(1)} = \gamma_{10}^{(1)}$ were rejected, with $\alpha = .05$ as the corresponding posterior probability p_0 was greater than .95. It follows that there is a main effect of ethnicity and of the clustering in colleges/universities.

TABLE 4

College Alcohol Problem Scale Randomized Response Data: Parameter Estimates of a Mixed and Fixed Effects Randomized Item Response Model

	Mixed Effects Model			Fixed Effects Model		
Parameter	*M*	*SD*	HPD	*M*	*SD*	HPD
Fixed effects						
$\gamma^{(2)}$	.118	.445	− 0.780, 0.969	.140	.157	− 0.179, 0.442
$\gamma^{(1)}$ (Gender)	− .261	.102	− 0.465, − 0.065	− .264	.109	− 0.483, − 0.052
Ethnicity						
$\gamma_2^{(1)}$ (Asian)	− .180	.293	− 0.758, 0.375	− .198	.324	− 0.854, − 0.412
$\gamma_3^{(1)}$ (White)	.089	.118	− 0.141, 0.321	.085	.141	− 0.195, 0.357
$\gamma_4^{(1)}$ (Black)	− .474	.178	− 0.837, − 0.137	− .465	.186	− 0.824, − 0.095
$\gamma_5^{(1)}$ (Other)	.543	.247	0.079, 1.027	.587	.240	0.092, 1.050
University						
$\gamma_6^{(1)}$ (Elon)	.188	.127	− 0.073, 0.417	.041	.092	− 0.144, 0.217
$\gamma_7^{(1)}$ (University of North Carolina)	− .148	.137	− 0.439, 0.105	− .288	.122	− 0.529, − 0.054
$\gamma_8^{(1)}$ (Wake Forest)	− .014	.159	− 0.370, 0.262	− .150	.154	− 0.452, 0.149
$\gamma_9^{(1)}$ (Guilford)	.474	.149	0.143, 0.722	.406	.161	0.080, 0.712
Random effects						
σ_e^2	.913	.066	0.787, 1.049	.912	.065	0.789, 1.038
τ^2	.721	.773	0.106, 1.861			

Note: HPD = highest posterior density.

Concluding Remarks

In this article, a mixed effects randomized item response model was developed for analyzing binary or polytomous hierarchical structured RR data. The principal advantage of the proposed model is the ability to treat a variety of special problems in a unified framework. In general, the statistical inference of RR data can be improved and/or expanded by assuming an individual response model that specifies the relation between the randomized item response data and an individual underlying behavior or an individual true response probability. This allows the computation of individual estimates of a sensitive behavior or true response probability although the true individual item responses are masked. It is also shown that respondents and items can be calibrated with IRT in the presence of binary/polytomous RR and/or direct questioning data.

Functional forms for relations between covariates at different levels and individual behaviors/attitudes or individual true response probabilities can be specified. The variation among individual sensitive latent characteristics or among individual true response probabilities can be explained by background variables, or in specific, effects of background variables on a sensitive latent characteristic can be explored. The proposed model allows the computation of individual estimates and their relationships with background variables where the observations may be obtained in clusters. Note that the RR technique results in less underreporting of sensitive behavior (the CAPS data analysis showed an increased cooperation of respondents using the RR interviewing technique), and furthermore, traditional methods for analyzing RR data are restricted to estimates of population proportions.

Both simulation studies showed that the model parameters can be accurately estimated given DQ and/or RR observations. The RR technique requires larger sample sizes to obtain parameter estimates with the same precision as those obtained via direct questioning. In the unrelated question design, there is an efficiency loss due to observing responses to the unrelated question. This loss of efficiency can be improved using relevant prior information. For example, the first simulation study showed that in a RR sampling design a reduction in sampling error can be obtained by using relevant grouping structures and/or background variables when estimating the proportion of positive responses.

The model can be extended by specifying relationships at the item level, for example, between covariates and item parameters. A model can be specified on the item parameters to model item variation across groups (see e.g., De Boeck & Wilson, 2004). An interesting case is to explore variation in the true proportion of positive responses, π_{ijk}. For example, assume that the groups are randomly selected from a larger population and that 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. That is,

$$\Phi^{-1}(\pi_{ijk}) = \mu_k + \zeta_{jk} + \epsilon_{ijk}, \tag{34}$$

where μ_k is the general mean, considering all responses to item k, ζ_{jk} is the random group effect, $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma_\epsilon^2)$, and $\zeta_{jk} \sim \mathcal{N}\left(0, \sigma_\zeta^2\right)$. The heterogeneity across groups in the random effects or group specific response probabilities, causing responses from the same group to be associated, is reflected in Equation 34. The model in Equation 34 can be extended by introducing individual or group characteristics for explaining differences between the individual true response probabilities, or for increasing the accuracy of the corresponding estimates, given observed randomized responses.

Appendix A
Markov Chain Monte Carlo (MCMC) Implementation

In general, the MCMC implementation consists of three parts: (a) sampling augmented nonobserved true item responses, denoted as $\tilde{\mathbf{y}}$, given observed randomized item responses; (b) sampling latent continuous true item responses, denoted as $\mathbf{z}$, given the augmented nonobserved true item responses; and (c) sampling all model parameters given the augmented continuous true item responses.

Sampling $\tilde{\mathbf{y}}$

The sampling of the augmented true item responses is described for the related and unrelated randomized response sampling design and for binary as well as polytomous data. For the unrelated question design it is assumed that the response probability to the unrelated question, p_2, is known. In the other case, additional sampling steps are needed as described on page 9. Assume that the probability of a positive true response to item k of respondent ij, is given by π_{ijk}.

1. Binary randomized response data.
   - The related question design. Via a path diagram (Fox, 2005b), it can be seen that the true nonobserved item responses are Bernoulli distributed. That is,

$$\begin{aligned} \tilde{y}_{ijk}|y_{ijk} = 1, \pi_{ijk} &\sim \mathcal{B}\left(\lambda = \frac{p\pi_{ijk}}{p\pi_{ijk} + (1-p)(1-\pi_{ijk})}\right) \\ \tilde{y}_{ijk}|y_{ijk} = 0, \pi_{ijk} &\sim \mathcal{B}\left(\lambda = \frac{(1-p)\pi_{ijk}}{p(1-\pi_{ijk}) + (1-p)\pi_{ijk}}\right), \end{aligned} \tag{A1}$$

   where λ defines the success probability of the Bernoulli distribution.
   - The unrelated question design. The nonobserved true item responses are Bernoulli distributed:

$$\begin{aligned} \tilde{y}_{ijk} \mid y_{ijk} = 1, \pi_{ijk} &\sim \mathcal{B}\left(\lambda = \frac{\pi_{ijk}\left(p_1 + p_2(1-p_1)\right)}{p_1\pi_{ijk} + p_2(1-p_1)}\right) \\ \tilde{y}_{ijk} \mid y_{ijk} = 0, \pi_{ijk} &\sim \mathcal{B}\left(\lambda = \frac{\pi_{ijk}(1-p_1)(1-p_2)}{1 - \left(p_1\pi_{ijk} + p_2(1-p_1)\right)}\right). \end{aligned} \tag{A2}$$

2. Polytomous ordinal response data for the unrelated question design.

The latent random variable $\widetilde{y}_{ijk}$ given $y_{ijk} = c$ is multinomial distributed with cell probabilities:

$$\Delta(c) = \frac{\pi_{ijk}(c')p_1 I(c = c') + \pi_{ijk}(c')(1-p_1)p_2(c)}{\pi_{ijk}(c)p_1 + (1-p_1)p_2(c)}, \tag{A3}$$

for c and $c' = 1, \ldots, C_k$.

Sampling z

1. Single item response. For the binary case, the variable $\widetilde{y}_{ijk}$ is Bernoulli distributed with success probability

$$\pi_{ijk} = \Phi(\mathbf{w}_{ij}^t \boldsymbol{\gamma} + \mathbf{x}_{ij}^t \boldsymbol{\beta}_j) \tag{A4}$$

for $j = 1, \ldots, J$ and $i = 1, \ldots, n_j$ according to Equation 9. A latent variable z_{ijk} is introduced that follows a truncated normal distribution (e.g., Albert & Chib, 1993)

$$z_{ijk} \mid \widetilde{y}_{ijk}, \boldsymbol{\beta}_j, \boldsymbol{\gamma} \sim \mathcal{N}\left(\mathbf{w}_{ij}^t \boldsymbol{\gamma} + \mathbf{x}_{ij}^t \boldsymbol{\beta}_j, 1\right) \tag{A5}$$

with $\widetilde{y}_{ijk}$ the indicator of z_{ijk} being positive. In the same way, for multinomial response data, the variable $\widetilde{y}_{ijk}$ is multinomial distributed with cell probabilities

$$\pi_{ijk}(c) = \Phi(\mathbf{w}_{ij}^t \boldsymbol{\gamma} + \mathbf{x}_{ij}^t \boldsymbol{\beta}_j - \kappa_{k,(c-1)}) - \Phi(\mathbf{w}_{ij}^t \boldsymbol{\gamma} + \mathbf{x}_{ij}^t \boldsymbol{\beta}_j - \kappa_{k,c}), \tag{A6}$$

for $c = 1, \ldots, C_k$. Subsequently, a latent continuous random variable z_{ijk} is distributed according to Equation A5 with $\widetilde{y}_{ijk} = c$ if $\kappa_{k,(c-1)} \leq z_{ijk} \leq \kappa_{k,c}$. The probit link function can be replaced by a logistic link function.

2. Multiple item responses. To implement the Gibbs sampler for binary response data a latent continuous random variable z_{ijk} is defined such that

$$z_{ijk} \mid \widetilde{y}_{ijk}, \theta_{ij}, \boldsymbol{\xi}_k \sim \mathcal{N}\left(a_k \theta_{ij} - b_k, 1\right), \tag{A7}$$

with $\widetilde{y}_{ijk}$ the indicator of z_{ijk} being positive, for $j = 1, \ldots, J$, $i = 1, \ldots, n_j$, and $k = 1, \ldots, K$, according to Equation 12. For polytomous response data define

$$z_{ijk} \mid \widetilde{y}_{ijk}, \theta_{ij}, \boldsymbol{\xi}_k \sim \mathcal{N}\left(a_k \theta_{ij}, 1\right), \tag{A8}$$

with $\widetilde{y}_{ijk} = c$ if $\kappa_{k,(c-1)} \leq z_{ijk} \leq \kappa_{k,c}$, according to Equation 13. Again, the probit link function can be replaced by a logistic link function; see Patz and Junker (1999a, 1999b).

Sampling Parameters

- Draw $\boldsymbol{\xi}_k$ given $\widetilde{\mathbf{y}}, \mathbf{z}, \boldsymbol{\theta}, \mu_\xi, \Sigma_\xi$.
  For binary data:

$$\boldsymbol{\xi}_k \mid \mathbf{z}_k, \boldsymbol{\theta}, \mu_\xi, \Sigma_\xi \sim \mathcal{N}\left(\Omega\left(H^t \mathbf{z}_k + \Sigma_\xi^{-1}\mu_\xi\right), \Omega\right), \tag{A9}$$

where $\Omega^{-1} = (H^t H)^{-1} + \Sigma_\xi^{-1}$ and $H = [\boldsymbol{\theta}, -\mathbf{1}]$. Hyperprior parameter Σ_ξ is sampled from an inverse-Wishart distribution:

$$\Sigma_\xi \mid \boldsymbol{\xi} \sim Inv - Wishart\left(K + n_0, \left(\sum \xi_k \xi_k^t + \Lambda\right)^{-1}\right) \tag{A10}$$

For polytomous data:

$$a_k \mid \mathbf{z}_k, \boldsymbol{\theta} \sim \mathcal{N}\left(\hat{a}_k, \left(\boldsymbol{\theta}^t \boldsymbol{\theta}\right)^{-1}\right) p(a_k) \tag{A11}$$

where $\hat{a}_k$ is the usual least squares estimator following from the linear regression from $\mathbf{z}_k$ on θ. The threshold parameters, κ_k, are sampled using the Metropolis-Hastings algorithm. A candidate is sampled from a normal distribution,

$$\kappa_{k,c}^* \sim \mathcal{N}\left(\kappa_{k,c}^{(m)}, \sigma_{MH}^2\right), \tag{A12}$$

where $\kappa_{k,c}^{(m)}$ is the value of $\kappa_{k,c}$ in the mth iteration of the sampler. This new candidate is accepted with probability (Fox, 2005a)

$$\prod_{i|j} \frac{P\left(\widetilde{y}_{ijk} \mid \theta_{ij}, a_k, \boldsymbol{\kappa}_k^*\right)}{P\left(\widetilde{y}_{ijk} \mid \theta_{ij}, a_k, \boldsymbol{\kappa}_k\right)} \times \prod_{c=1}^{C_k - 1} \frac{\Phi(\kappa_{k,(c+1)} - \kappa_{k,c})/\sigma_{MH} - \Phi(\kappa_{k,(c-1)}^* - \kappa_{k,c})/\sigma_{MH}}{\Phi(\kappa_{k,(c+1)}^* - \kappa_{k,c}^*)/\sigma_{MH} - \Phi(\kappa_{k,(c-1)} - \kappa_{k,c}^*)/\sigma_{MH}}. \tag{A13}$$

- Draw $\boldsymbol{\theta}$ given $\mathbf{z}, \boldsymbol{\xi}_k, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_e^2$.
  The full conditional of $\boldsymbol{\theta}$ is specified by the mixed effects model, Equation 15, and the linear regression of $\mathbf{z}$ on $\boldsymbol{\theta}$, with item parameters $\boldsymbol{\xi}$ as regression coefficients. It follows that the full conditional is normally distributed:

$$\theta_{ij} \mid \mathbf{z}_{ij}, \boldsymbol{\xi}, \boldsymbol{\beta}_j, \boldsymbol{\gamma}, \sigma_e^2 \sim \mathcal{N}\left(\frac{\widehat{\theta}_{ij}/\nu + (\mathbf{w}_{ij}^t \boldsymbol{\gamma} + \mathbf{x}_{ij}^t \boldsymbol{\beta}_j)/\sigma_e^2}{\nu^{-1} + \sigma_e^{-2}}, \frac{1}{\nu^{-1} + \sigma_e^{-2}}\right), \tag{A14}$$

where $\nu = \left(\sum_{k=1}^K a_k^2\right)^{-1}$ and $\widehat{\theta}_{ij}$ the least squares estimator following from the regression of $\mathbf{z}_{ij} + \mathbf{b}$ on $\mathbf{a}$ for binary data and $\mathbf{z}_{ij}$ on $\mathbf{a}$ for polytomous data.

- Draw $\boldsymbol{\beta}$ given $\boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma_e^2, T$.
  The full conditional of each random effect is a multivariate normal distribution (e.g., Zeger & Karim, 1991)

$$\boldsymbol{\beta}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\gamma}, \sigma_e^2, T \sim \mathcal{N}\left(\mathbf{D}\mathbf{x}_j^t(\boldsymbol{\theta}_j - \mathbf{w}_j\boldsymbol{\gamma})/\sigma_e^2, \mathbf{D}\right). \tag{A15}$$

where

$$\mathbf{D}^{-1} = \mathbf{x}_j^t \mathbf{x}_j / \sigma_e^2 + T^{-1}.$$

- Draw $\boldsymbol{\gamma}$ given $\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_e^2, T$.
  The full conditional of the fixed effects is a multivariate normal distribution:

$$\boldsymbol{\gamma} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_e^2, T \sim \mathcal{N}\left(\left(\mathbf{w}^t\mathbf{w} + \sigma_\gamma^{-1}\sigma_e^2 I_p\right)^{-1} \mathbf{w}^t(\boldsymbol{\theta} - \mathbf{x}\boldsymbol{\beta}), \left(\mathbf{w}^t\mathbf{w} + \sigma_\gamma^{-1}\sigma_e^2 I_p\right)^{-1} \right). \qquad \text{(A16)}$$

- Draw σ_e^2 given $\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}$.
  The full conditional of variance parameter σ_e^2 is an inverse gamma distribution with parameter $(n_1 + N)/2$ and scale parameter $(s_1 + \sum_{i|j}(\theta_{ij} - (\mathbf{w}_{ij}\boldsymbol{\gamma} + \mathbf{x}_{ij}\boldsymbol{\beta}_j))^2)/2$.
- Draw T given $\boldsymbol{\beta}$.
  The full conditionals of covariance matrix T is an inverse-Wishart distribution with degrees of freedom $n_q + J$ and scale parameter $S + \sum_j \boldsymbol{\beta}_j \boldsymbol{\beta}_j^t$.

Monitoring Convergence

Initial values for the sampler can be obtained by fitting a mixed effects model to the data but ignoring the randomized response character of the discrete data. The mixed effects model parameters will be biased but can serve as starting values for the MCMC algorithm. To determine whether the algorithm has converged, the multivariate version of the statistic of Brooks and Gelman (1998) can be used. This requires s multiple runs of the sampler from overdispersed starting values. Furthermore, the convergence of the MCMC chains can be checked using the standard convergence diagnostics implemented in the BOA software (http://www.public-health.uiowa.edu/boa). In addition, plots of the average of each parameter across multiple chains, and the running average can provide extra information.

Appendix B
College Alcohol Problem Scale Questionnaire

How often (almost always [5], often [4], sometimes [3], seldom [2], almost never [1]) have you had any of the following problems over the past years as a result of drinking too much alcohol?

1. Feeling sad, blue or depressed
2. Nervousness or irritability
3. Hurt another person emotionally
4. Family problems related to your drinking
5. Spent too much money on drugs
6. Badly affected friendship or relationship
7. Hurt another person physically
8. Caused other to criticize your behavior
9. Nausea or vomiting
10. Drove under the influence

11. Spent too much money on alcohol
12. Feeling tired or hung over
13. Illegal activities associated with drug use

References

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*, 669–679.

Böckenholt, U., & van der Heijden, P. G. M. (2007). Item-randomized response models for measuring noncompliance: Risk-return perceptions, social influences, and self-protective responses. *Psychometrika*, *72*, 245–262.

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

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455.

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

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach* (Statistics for Social Science and Behavorial Sciences). New York: Springer.

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

Fox, J.-P. (2005a). Multilevel IRT using dichotomous and polytomous items. *British Journal of Mathematical and Statistical Psychology*, *58*, 145–172.

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

Fox, J.-P. (in press). Beta-binomial ANOVA for multivariate randomized response data. *British Journal of Mathematical and Statistical Psychology*.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

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.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). New York: Oxford University Press.

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.

Hedeker, D. R., & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, *50*, 933–944.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*, 963–974.

Lensvelt, G. J. L. M., Hox, J. J., van der Heijden, P. G. M., & Maas, C. (2005). Meta-analysis of randomized response research: 35 years of validation. *Sociological Methods and Research*, *33*, 319–348.

Lensvelt, G. J. L. M., van der Heijden, P. G. M., & Laudy, O. (2006). A validation of a computer-assisted randomized response survey to estimate the prevalence of fraud in social security. *Journal of the Royal Statistical Society A*, *169*, 305–318.

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced linear mixed models with nested random effects. *Biometrika*, *74*, 817–827.

Maddala, G. S. (1983). *Limited dependent and qualitative variables in econometrics*. Cambridge, UK: Cambridge University Press.

O'Hare, T. M. (1997). Measuring problem drinkers in first time offenders: Development and validation of the College Alcohol Problem Scale (CAPS). *Journal of Substance Abuse Treatment*, *14*, 383–387.

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.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs*, *34*(Suppl. 17).

Scheers, N. J., & Dayton, C. (1988). Covariate randomized response model. *Journal of the American Statistical Association*, *83*, 969–974.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. New York: Chapman & Hall/CRC.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis*. London: Sage.

Tracy, P., & Fox, J. (1981). The validity of randomized response for sensitive measurements. *American Sociological Review*, *46*, 187–199.

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

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.

Authors

J.-P. FOX is assistant professor, Department of Research Methodology, Measurement, and Data Analysis, the University of Twente, Enschede, P.O. Box 217, 7500 AE Enschede, Netherlands; e-mail: J.P.Fox@utwente.nl. His areas of specialization include developing and applying Bayesian techniques and statistical models in test theory and other scientific disciplines.

CHERYL WYRICK is research associate, Tanglewood Research, 420 Gallimore Dairy Rd. Suite A, Greensboro, NC 27409; e-mail: chwyrick@uncg.edu. Her areas of specialization include survey research methods and program evaluation as they relate to substance abuse prevention.

Manuscript received July 18, 2006
Accepted April 25, 2007