Psychological Test and Assessment Modeling, Volume 57, 2015 (4), 523-541

Nonignorable data in IRT models: Polytomous responses and response propensity models with covariates

C. A. W. Glas[1], *J. L. Pimentel*[2] *& S. M. A. Lamers*[3]

Abstract

Missing data usually present special problems for statistical analyses, especially when the data are not missing at random, that is, when the ignorability principle defined by Rubin (1976) does not hold. Recently, a substantial number of articles have been published on model-based procedures to handle nonignorable missing data due to item nonresponse (Holman & Glas, 2005; Glas & Pimentel, 2008; Rose, von Davier & Xu, 2010; Pohl, Gräfe & Rose, 2014). In this approach, an item response theory (IRT) model for the observed data is estimated concurrently with an IRT model for the propensity of the missing data.

The present article elaborates on this approach in two directions. Firstly, the preceding articles only consider dichotomously scored items; in the present article it is shown that the approach equally works for polytomously scored items. Secondly, it is shown that the methods can be generalized to allow for covariates in the model for the missing data. Simulation studies are presented to illustrate the efficiency of the proposed methods.

Keywords: item response theory, latent traits, missing data, nonignorable missing data, observed covariates

[1] *Correspondence concerning this article should be addressed to:* C. A. W. Glas, PhD, University of Twente, Cubicus C338, The Netherlands; email: c.a.w.glas@utwente.nl

[2] Mindanao State University, Iligan Institute of Technology, the Philippines

[3] University of Twente, the Netherlands

Introduction

Missing data are always a source of concern for statistical analyses. It raises the level complexity of statistical inference. Many researchers, methodologists, and software developers resort to editing the data, although ad hoc edits may do more harm than good by producing results that are substantially biased, inefficient and unreliable (Schafer & Graham, 2002). One way to address the bias in parameter estimates is the identification of the variables that explain the cause of missing data. Below, these explanatory variables will be called "mechanism or process" variables. By including a model for this missing data mechanism in the estimation we can reduce or eliminate the bias in parameter estimates.

Theoretically, if all the process variables associated with a particular piece of missing data can be identified and modeled accurately as controls, the impact of the missing data can be statistically adjusted to the point where it is ignorable (Little & Rubin, 1987). In practice, it is difficult to identify these process variables for all cases of missing data. However, if the given data set contains missing observations, the mechanism causing this missingness can be characterized by its variety of randomness (Rubin, 1976) as missing completely at random (MCAR), missing at random (MAR) and not missing at random (NONMAR).

In this article, we focus on responses of persons to items and on item nonresponse. Suppose θ and ζ are the parameters of the observed data and the missing data process, respectively, and D is the missing data indicator with elements $d_{ik} = 1$ if a realization x_{ik} was observed and $d_{ik} = 0$ if x_{ik} was missing for persons i and items k. Following Rubin's definition, missing data is MAR if the probability of D given the observed data x_{obs}, missing data x_{mis}, and observed covariates y does not depend on the missing data x_{mis}, that is, if

$$p(D \mid x_{obs}, x_{mis}, \zeta, y) = p(D \mid x_{obs}, \zeta, y). \tag{1}$$

Furthermore, the parameters θ and ζ are distinct if there are no functional dependencies between them, that is, restrictions on the parameter space (frequentist version) or if the prior distributions of ζ and θ are independent (Bayesian case). (It should be noted that this is a somewhat rough definition, for technical details refer to Rubin (1976), Heitjan (1994, 1997), Heitjan and Rubin (1991), and Jaeger (2005)). If MAR and distinctness hold, the missing data is said to be ignorable, otherwise the missing data are nonignorable. If ignorability holds, we do not have to take the distribution of D and ζ into account, and the consistency of the estimates is not threatened by the occurrence of the missing data.

In the framework of IRT, missing data can be split into four types (Lord, 1974). The first consists of missing observations which result from a priori fixed incomplete test administration and calibration designs. In this case, the missing data are a priori fixed and ignorability trivially holds. That is, $p(D \mid x_{obs}, x_{mis}, \zeta, y) = p(D \mid x_{obs}, \zeta, y) = 1$. The second type consists of classes of response-contingent designs such as two-stage and multistage

testing designs and computerized adaptive testing (Lord, 1980). These designs produce ignorable missing data, because the design variables *D* are completely determined by the observed responses (see, for instance, Mislevy & Wu, 1996). The third type is ignorable missing data that results from unscalable responses such as items missing from booklets and responses such as "do not know" or "not applicable"'. Missing pages can be reasonably viewed as missing at random; "do not know"' or "not applicable" are already suspicious and might fall in the next category of missing data. The reason is that it is not automatically clear whether the given response (don't know or not applicable) is accurate or an instance of avoidance behavior. The fourth and last type of missing data results from a nonignorable missing data mechanism. These will, for instance, occur when low-ability respondents fail to give responses to specific items as a result of discomfort or embarrassment. In the framework of a medical survey, Holman and Glas (2005) report that patients with a relatively high functional status boosted the estimate of their ability level by failing to respond to items of a physical disability scale. Of course, whether ignorability holds or not, needs to be tested in these cases.

Moustaki (1996, see also Bartholomew & Knott, 1999) developed a general latent trait and latent class model for mixed observed variables which applies to Lord's fourth type of missing data. Three methods for dealing with nonignorable missing discrete data were proposed (O'Muircheartaigh & Moustaki, 1999; Moustaki & O'Muircheartaigh, 2000; Moustaki & Knott, 2000). In the first method for the treatment of nonresponse, the missing value is treated as a separate response category. So the method includes the missing values in the analysis of the observed items. That is, it is assumed that responses and nonresponses are related to the same attitude dimension or dimensions.

The second method to deal with nonresponse is computing response propensities. The idea is to use a propensity score to weight item responses and respondents to account for item and unit nonresponse and to obtain adjusted estimates. This response propensity method uses a logistic or probit regression model which is fitted to a binary item response-nonresponse variable for the item of interest with a set of covariates. The third method is to use a latent variable model with two latent dimensions, one to summarize the response propensity and the other to summarize the individual position on the dimension of interest (such as ability or attitude). As an example, O'Muircheartaigh and Moustaki (1999) used a latent variable model for the treatment of item nonresponse in attitude scales. This latent variable approach allows missing values to be included in the analysis and, equally important, allows information about attitude to be inferred from nonresponse. Their method handles binary (dichotomous), metric and mixed (binary and metric) manifest items with missing values. The second and third methods are closely related: both entail the estimation of a response propensity distribution and can be seen as an elaboration of methods of adjustment by propensity scoring proposed by Heckman (1979).

Working within the third approach, Holman and Glas (2005) proposed an IRT model for skipped items that allows concurrent estimation of IRT item parameters for both a model for the observed dichotomous responses and the missing data indicators. Other applications pertain to items that are not reached, such as missing item responses at the end of a speeded test (Glas & Pimentel, 2008). Rose (2013) generalized these approaches to allow

for multidimensional IRT models for the missing data indicators and showed that disregarding the multidimensional latent structure can lead to a failure to correct for nonignorable item nonresponse. Further, Rose (2013) considered an alternative approach where the average number of missing responses is used as a covariate in the distribution of the latent variables of an IRT model for the observed responses. This approach has both advantages (works well with small sample sizes and small numbers of missing responses) and disadvantages (attenuated estimates when the reliability of the covariate is low). Finally, Pohl, et al. (2014) studied a version of the model with a combination of propensity distributions for omitted and not-reached items. Also, they presented examples of empirical applications where a propensity distribution was not needed to correct for bias in parameter estimates. In such cases, the actually observed responses contain enough information to account for missing responses.

In this article, we generalize model-based adjustment using IRT in two directions: IRT models for polytomously scored items and covariates for the propensity distribution. This article consists of five sections and is organized in the following manner. After this introduction, the model will be outlined and a maximum marginal likelihood (MML) estimation procedure will be sketched. Then some simulation studies will be presented to illustrate the performance of the procedure. The next section presents an empirical simulation, that is, a simulation using real data, to give an example of how the model can work in practice. The article concludes with a discussion.

IRT models

General IRT model for missing data

Suppose, for respondents labeled $i = 1,...,N$ and items labeled $k = 1,...,K$, that $x_{ik} \in \{0.,1,2,...,m\}$ is the observed item response and let $p(x_{ik} \mid d_{ik} = 1, \theta_i, \alpha_k, \beta_k)$ be the IRT model for the observed item response with item parameters α_k and β_k and person parameters θ_i. If $d_{ik} = 0$, we assume that x_{ik} is equal to an arbitrary constant c and so $p(x_{ik} = c \mid d_{ik} = 0, \theta_i, \alpha_k, \beta_k) = 1$. Let $p(d_{ik} \mid \zeta_i, \gamma_k, \delta_k)$ be the IRT model for the missing data indicator defined above, with item parameters γ_k and δ_k and person parameters ζ_i. All parameters may be vector-valued. It is assumed that the person's latent variables have a multivariate normal distribution with density $g(\zeta_i, \theta_i \mid \Sigma, H, y_i)$, where y_i are observed covariates, H are regression coefficients for the regression of $\lambda_i = (\theta_i, \gamma_i)$ on y_i, and Σ is the covariance matrix of the residuals. Note that if the covariates are lacking, Σ reduces to the covariance matrix of the latent person variables. Usually, Σ is restricted to a correlation matrix to identify the model. The likelihood of the model is expressed as

$$\prod_{i=1}^{N}\prod_{k=1}^{K} p(x_{ik} \mid d_{ik}, \theta_i, \alpha_k, \beta_k)\, p(d_{ik} \mid \zeta_i, \gamma_k, \delta_k)\, g(\zeta_i, \theta_i \mid \Sigma, H, y_i) \quad . \tag{2}$$

Note that by the definition of $p(x_{ik} \mid d_{ik}, \theta_i, \alpha_k, \beta_k)$, the unobserved values of x_{ik} are ignored in the likelihood given by (2). The idea behind the method is that ignorability is violated when the covariance between θ and ζ is non-zero. In these cases, the estimation procedure must be based on the complete model, so including the model for the response propensities.

Next, we will specify the three densities in formula (2). The propensity distribution $p(d_{ik} \mid \zeta_i, \gamma_k, \delta_k)$ must be chosen to reflect the process that caused the missing responses. For instance, Korobko, et al. (2008) choose an IRT model with single peaked response curves to model the choice of examination topics. In the present application, we assume the simpler situation of skipped items, and choose a propensity model that is a multidimensional generalization of the model by Holman and Glas (2005). So also "not reached" is not considered here. To model the missing data process, we use a Q_1-dimensional IRT model proposed by Reckase (1985, 1997) and Ackerman (1996a & 1996b). The probability of an observation is given by

$$p(d_{ik} = 1 \mid \zeta_i, \gamma_k, \delta_k) = \frac{\exp\left(\sum_q^{Q_1} \gamma_{kq}\zeta_{iq} - \delta_k\right)}{1 + \exp\left(\sum_q^{Q_1} \gamma_{kq}\zeta_{iq} - \delta_k\right)} . \qquad (3)$$

The model becomes the two parameter logistic (2PL) model (Lord & Novick, 1968) when $Q_1 = 1$ and the Rasch model (Rasch, 1960) when, in addition, $\gamma_{kq} = 1$.

For the observed responses, we consider the multidimensional generalized partial credit model (GPCM; Muraki, 1992). The probability of responding in a category g of item k by person i is given by

$$p(x_{ik} = g \mid d_{ik} = 1, \theta_i, \alpha_k, \beta_k) = \frac{\exp\left(g\sum_{q=1}^{Q_2} \alpha_{kq}\theta_{iq} - \beta_{kg}\right)}{1 + \sum_{h=1}^{m} \exp\left(h\sum_{q=1}^{Q_2} \alpha_{kq}\theta_{iq} - \beta_{kh}\right)} . \qquad (4)$$

Note that the model has a Q_2-dimensional latent person parameter $\theta_i = (\theta_{i1}, ..., \theta_{iq}, ..., \theta_{iQ_2})$, $\alpha_k = (\alpha_{k1}, ..., \alpha_{kq}, ..., \alpha_{kQ_2})$ are discrimination parameters and $\beta_k = (\beta_{k1}, ..., \beta_{kq}, ..., \beta_{kQ_2})$ are location parameters for the item response categories.

The latent person parameters λ_i (defined above as the concatenation of θ_i and ζ_i are assumed to have a Q-variate normal distribution ($Q = Q_1 + Q_2$), that is,

$$g(\lambda_i \mid \Sigma, H, y_i) = (2\pi)^{-Q/2} \mid \Sigma \mid^{-1/2} \exp\left(-1/2(\lambda_i - H^t y_i)^t \Sigma^{-1}(\lambda_i - H^t y_i)\right), \qquad (5)$$

where H is a P X Q matrix of regression coefficients and Σ is a Q X Q variance-covariance matrix. Equivalently, the general model for the latent variables can be expressed in matrix form as

$$\lambda = YH + E, \tag{6}$$

where λ is the N X Q matrix of latent variables, Y is a N X P matrix of observed covariates, and E is the N X Q matrix of residuals.

MML estimation

In the present article, the parameters of the model are estimated by maximum marginal likelihood (MML, see, Bock, Gibbons & Muraki, 1988). The likelihood given by formula (2) is marginalized by integrating over the person parameters, such that the marginal likelihood given by

$$\prod_{i=1}^{N} \int \cdots \int \prod_{k=1}^{k} p(x_{ik} \mid d_{ik}, \theta_i, \alpha_k, \beta_k)\, p(d_{ik} \mid \zeta_i, \gamma_k, \delta_k)\, g(\zeta_i, \theta_i \mid \Sigma, H, y_i)\, d\theta_i d\zeta_i \tag{7}$$

is a function of item parameters $\alpha_k, \beta_k, \gamma_k, \delta_k$, regression parameters H and the covariance matrix Σ only. Glas (1992, 1999, 2010a) shows that the likelihood equations can easily be derived using Fisher's identity (Efron, 1977; Louis 1982). The procedure boils down to deriving the likelihood equations assuming λ known, and then taking the posterior expectation of both sides of the equation. For instance, applied to the regression parameters, the likelihood equations assuming λ known are given by

$$\widehat{H} = (Y^t Y)^{-1} Y^t \lambda \tag{8}$$

and taking posterior expectations results in

$$\widehat{H} = (Y^t Y)^{-1} \sum_{i=1}^{N} y_i\, E(\lambda_i \mid x_i, d_i, y_i)^t, \tag{9}$$

where the Q-dimensional column-vector $E(\lambda_i \mid x_i, d_i, y_i)$ is the posterior distribution of λ_i given all observations on respondent i. Estimation equations for the other parameters are derived analogously (refer to Glas 1992, 1999, 2010a, also see, Adams, Wilson, & Wang, 1997, for an alternative derivation).

Simulation studies

Simulation studies were undertaken to assess the effect of the presence of nonignorable missing data on the estimates of item parameters and the effectiveness of the proposed methods to improve the estimates. It must be noted in advance that these simulations did not have the pretention of being exhaustive, because there are too many possible configurations of multidimensional IRT models for polytomously scored items and possible

propensity models to meet such a goal. So the primary aim was to check whether the approach behaved as expected in a limited set of conditions.

The simulation study consisted of two parts. The first part extended the study by Holman and Glas (2005) to a situation where the model for the missing data indicators is multidimensional, and studied the effects of including no, part of, or all latent dimensions of this model in the estimation. The motivation for this part of the study was that Rose (2013) showed that ignoring the multidimensional structure of the response propensity model can lead to substantial bias. In the present study, it was investigated whether this phenomenon also appeared here. The second simulation study pertained to the effects of adding observed covariates to the model. All simulations reported in this article were conducted with the public domain program MIRT (Glas, 2010b) using an MML estimation procedure.

Multidimensional response propensity models

To study the effects of including no, part of, or all latent dimensions of the missing data process in the estimation procedure, latent person parameters were drawn from a three-variate normal distribution. The sample size was $N = 500$ persons. The variance of the latent variables was always equal to one. The correlation between the latent trait variables θ_i and ζ_i, say, $\rho(\theta,\zeta)$, was varied as 0.0, 0.4 and 0.8. The values 0.0, 0.4 and 0.8 are chosen to clarify the effect of no, a small and a large violation of ignorability. The simulations by Holman and Glas (2005) showed that violations of ignorability related to latent correlations less than 0.4 had little impact and that the effects became manifest starting from a value of 0.4. Models with $\rho(\theta,\zeta)=0.0$ will be referred to as MAR models because they assume that the covariance between the latent variables pertaining to the item responses and the latent variables pertaining to the response propensities are zero. Other models will be called NONMAR models for the analogous opposite reason. Also the correlations between the two dimensions of the missing data process $\rho(\zeta_1,\ \zeta_2)$ were varied as 0.0, 0.4 and 0.8. The items were either dichotomously and polytomously scored. The test consisted of $K = 10$ items. The values x_{ik} and d_{ik} were drawn from $p(x_{ik}\,|\,d_{ik},\theta_i,\alpha_k,\beta_k)$ and $p(d_{ik}\mid\ \zeta_i,\gamma_k,\delta_k)$, respectively. The data were used to compute MML estimates of the item parameters under the various assumptions. Then the values of item parameters estimates over replications r *(r=1,...,R,* with *R=100)*, say $\hat{\phi}_r$, were compared with the values of the parameters used to generate the data using the mean absolute error, that is

$$MAE=\frac{1}{R}\sum_{r=1}^{R}\left|\hat{\phi}_r-\phi\right| \ . \tag{10}$$

For the dichotomous case, two conditions were used. In the first condition, the item parameters for all items were $\alpha_k=\gamma_k=1.0$, $\beta_k=0.0$, and $\delta_k=-1.0$. These values resulted in about 25% missing data. In the second condition, we used $\alpha_k=\gamma_k=1.0$ and, $\beta_k=\delta_k=0.0$, which resulted in about 50% missing data. The motive for fixing the item parameters was that it leads to a clear interpretation of the MAE relative to the fixed

value. Not reported here are simulation studies carried out where the item parameters were drawn from normal distributions with the reported fixed values as expectation and standard deviations of 0.5. They produced results similar to the results reported below. The MAE of the item parameter estimates are given in Table 1. For the polytomous case, items with three response categories were used in the simulation. The location parameters of the IRT model for the response propensities were varied as $\delta_k = 0.0$, and $\delta_k = -1.0$ for all k. The other item parameters were the same across conditions, that is, $\alpha_k = \gamma_k = 1.0$, $\beta_{k1} = -1.0$, and $\beta_{k2} = 1.0$, for all k. The MAEs of the parameter estimates are given in Table 2.

Both tables have the same format. The first column, labeled δ, refers to the location parameter for the IRT model for the response propensities used for generating the data. The first row pertains to a baseline condition where $\rho(\theta,\zeta) = 0.0$. So there ignorability holds. The values of the *MAE(α)* and *MAE(β)* are given in the two columns labeled α and β; they are the mean absolute errors over the 100 replications, and they serve as a baseline. The next three rows pertain to data generated using $\rho(\theta,\zeta) = 0.4$ and $\rho(\zeta_1, \zeta_2) = 0.0$. These data were analyzed using no, one and two dimensions for the missing data indicator, as indicated in the column labeled Q_2. The lines below, give the results for further values of $\rho(\theta,\zeta)$, $\rho(\theta_1,\theta_2)$, and Q_2. In Table 1, the columns labeled α, β, δ and γ give the MAEs for the respective item parameters. For polytomous case, reported in Table 2, there are two columns for the mean absolute error of the location parameters, referred as β_1 and β_2.

The simulations showed that the MAE values of the item parameter estimates were inflated when no model for missing data process was used in the parameter estimation. The effect increased as correlation between the latent variables for both observed data and the missing data process increased. For instance, if we consider Table 1, when $\delta = 0$, (that is, when there were 50% missing data) the baseline simulation of MAR data resulted in *MAE(α) = 0.225* and *MAE(β) = 0.120*. When $\rho(\theta,\zeta) = 0.4$ and $\rho(\zeta_1, \zeta_2) = 0.0$, and the missing data were ignored ($Q_2 = 0$), the MAE for α and β had values 0.245 and 0.148, respectively. This main effect was generally present both for dichotomously and polytomously scored items. So the first conclusion is that ignoring the missing data process lead to inflated estimation errors.

When the model for the missing data process was included in the analysis, that is, when the NONMAR model was used, the MAE values dropped to 0.228 for α and 0.133 for β when $Q_2 = 1$, and to *MAE(α) = 0.223* and *MAE(β) = 0.128* when $Q_2 = 2$. In general, a decrease in the values of the MAE of the item parameters was observed and this decrease was positively related to the number of dimensions included. Similar results were also observed for the values of MAE of the item parameters δ and γ for missing data process. So the second conclusion is that invoking the missing data process leads to a reduction of estimation errors, even if not all dimensions are invoked.

The third conclusion that can be drawn from the tables is that when the missing data process was completely modeled, that is, when $Q_2 = 2$, the estimation errors could even fall below the errors of the baseline. For instance, in Table 1 we see that for $\delta = 0$, the

Table 1:
MAE of item parameter estimates under MAR and NONMAR model;
Dichotomously scored items

Generating Values			Analysis	Mean Absolute Error			
δ	$\rho(\theta, \zeta)$	$\rho(\zeta_1, \zeta_2)$	Q_2	α	β	δ	γ
-1.0	0.0	-	-	0.168	0.102		
	0.4	0.0	0	0.169	0.118		
			1	0.163	0.113	0.126	0.467
			2	0.163	0.113	0.106	0.162
		0.4	0	0.169	0.107		
			1	0.164	0.102	0.109	0.205
			2	0.165	0.102	0.108	0.149
		0.8	0	0.170	0.110		
			1	0.165	0.104	0.100	0.137
			2	0.165	0.104	0.104	0.140
	0.8	0.4	0	0.176	0.140		
			1	0.156	0.105	0.101	0.151
			2	0.160	0.104	0.099	0.135
		0.8	0	0.170	0.137		
			1	0.153	0.103	0.099	0.136
			2	0.156	0.103	0.103	0.137
0.0	0.0	-	-	0.225	0.120		
	0.4	0.0	0	0.245	0.148		
			1	0.228	0.133	0.081	0.568
			2	0.223	0.128	0.089	0.154
		0.4	0	0.229	0.142		
			1	0.209	0.124	0.079	0.194
			2	0.209	0.125	0.084	0.147
		0.8	0	0.222	0.144		
			1	0.214	0.121	0.086	0.126
			2	0.214	0.122	0.088	0.126
	0.8	0.4	0	0.257	0.210		
			1	0.187	0.128	0.078	0.158
			2	0.186	0.129	0.083	0.136
		0.8	0	0.245	0.220		
			1	0.192	0.133	0.083	0.121
			2	0.194	0.133	0.083	0.121

Table 2:
MAE of item parameter estimates under MAR and NONMAR model;
Polytomously scored items

Generating Values			Analysis	Mean Absolute Error				
δ	$\rho(\theta, \zeta)$	$\rho(\zeta_1, \zeta_2)$	Q_2	α	β_1	β_2	δ	γ
-1.0	0.0	-	-	0.137	0.135	0.194		
	0.4	0.0	0	0.142	0.129	0.194		
			1	0.139	0.126	0.192	0.127	0.470
			2	0.138	0.125	0.193	0.103	0.167
		0.4	0	0.138	0.139	0.198		
			1	0.136	0.129	0.194	0.109	0.192
			2	0.135	0.129	0.194	0.107	0.162
		0.8	0	0.136	0.137	0.206		
			1	0.133	0.126	0.193	0.098	0.138
			2	0.132	0.127	0.192	0.100	0.139
	0.8	0.4	0	0.152	0.153	0.247		
			1	0.140	0.129	0.200	0.100	0.154
			2	0.138	0.126	0.200	0.103	0.138
		0.8	0	0.138	0.150	0.241		
			1	0.129	0.125	0.196	0.098	0.130
			2	0.128	0.125	0.197	0.102	0.130
0.0	0.0	-	-	0.187	0.156	0.239		
	0.4	0.0	0	0.182	0.173	0.259		
			1	0.175	0.148	0.250	0.080	0.548
			2	0.174	0.145	0.242	0.088	0.155
		0.4	0	0.189	0.173	0.257		
			1	0.182	0.150	0.241	0.078	0.182
			2	0.182	0.150	0.243	0.084	0.143
		0.8	0	0.188	0.182	0.274		
			1	0.183	0.151	0.246	0.087	0.131
			2	0.183	0.151	0.246	0.090	0.129
	0.8	0.4	0	0.197	0.228	0.367		
			1	0.165	0.148	0.245	0.081	0.152
			2	0.167	0.143	0.247	0.086	0.137
		0.8	0	0.195	0.241	0.411		
			1	0.170	0.154	0.250	0.088	0.120
			2	0.171	0.153	0.249	0.090	0.123

MAE(α) = 0.225 for the baseline and *MAE(α) = 0.194* for $\rho(\theta,\zeta) = \rho(\zeta_1, \zeta_2) = 0.8$ and $Q_2 = 2$. The tentative conclusion is that invoking a model for the missing data indicator resulted in additional collateral information which improved the estimates.

The fourth conclusion pertains to the extent to which MAR was violated. Inspection of the tables shows that if we ignored the missing data process ($Q_1 = 0$), the magnitude of the estimation error for $\rho(\theta,\zeta) = 0.8$ was greater than the magnitude for $\rho(\theta,\zeta) = 0.4$. For instance, in Table 1 we see that conditionally on $\rho(\zeta_1, \zeta_2) = 0.4$, the MAEs for α were 0.229 and 0.257, respectively. Finally, there was no clear effect of $\rho(\zeta_1, \zeta_2)$. That is, the effect was clear insofar that for the item parameters of the latent ability, the actual dimensionality of the missing propensity was irrelevant.

Response propensity models with observed covariates

The simulation procedure used was analogous to the simulation procedure in the previous section, but with the added feature of including observed covariates. To achieve comparability with the previous section, the regression coefficients were chosen as follows. Let Σ_λ be the covariance matrix of both the latent variables for the observed responses and the missing data indicator. As in the previous section, there was one dimension for the observed responses and there were two dimensions for the missing data process. Only the case $\rho(\zeta_1, \zeta_2) = 0.8$ was considered here. Further, either $\rho(\theta,\zeta) = 0.4$ or $\rho(\theta,\zeta) = 0.8$. Let Σ_ε be the diagonal matrix of the variances of the error terms. These variances were all equal to 0.15. For the data generation, the variables y_i had Q-variate independent standard normal distributions, and the regression coefficients H were chosen such that

$$\Sigma_\lambda = HH^t + \Sigma_\varepsilon \ , \tag{11}$$

where H is a lower-triangular matrix. As before, the sample size was $N = 500$. Again the test length was $K = 10$ and the item parameters were also as used above. One hundred replications were made for every combination of δ, $\rho(\theta,\zeta)$ and Q_2.

The results are given in the tables 3 and 4. The format of the tables is analogous to the previous two tables, except for an added column *P*, which refers to the number of covariates included in the parameter estimation. Note that also the baseline model where $\rho(\theta,\zeta) = 0.0$ (the MAR model) includes a covariate. This was done to enable the comparison with the NONMAR models.

Referring to Table 3, the baseline MAR model resulted in *MAE(α) = 0.145* and *MAE(β) = 0.117* for the case $\delta = 0$, i.e., 50% missing data, as compared to the analogous simulation reported in Table 1, where *MAE(α) = 0.225* and *MAE(β) = 0.120*. This increase in precision is due to the inclusion of a covariate. When we increased the correlation to $\rho(\theta,\zeta) = 0.4$ and $\rho(\theta,\zeta) = 0.8$, results showed that when the missing data process was ignored and only the covariate for θ was included, the values *MAE(α) = 0.170* and *MAE(β) = 0.164* were obtained. When a response propensity model with one latent di-

Table 3:
MAE of item parameter estimates under MAR and NONMAR model with covariates. Dichotomously scored items

Generating Values			Analysis		Mean Absolute Error			
δ	$\rho(\theta, \zeta)$	$\rho(\zeta_1, \zeta_2)$	Q_2	P	α	β	δ	γ
-1.0	0.0	-	-	1	0.121	0.095		
	0.4	0.8	0	1	0.157	0.117		
			1	2	0.131	0.097	0.102	0.150
			2	3	0.120	0.097	0.095	0.104
	0.8	0.8	0	1	0.188	0.142		
			1	2	0.159	0.110	0.092	0.127
			2	3	0.126	0.101	0.089	0.100
0.0	0.0	-	-	1	0.145	0.117		
	0.4	0.8	0	1	0.170	0.164		
			1	2	0.150	0.116	0.083	0.134
			2	3	0.140	0.114	0.078	0.094
	0.8	0.8	0	1	0.313	0.228		
			1	2	0.204	0.134	0.092	0.125
			2	3	0.155	0.126	0.084	0.097

Table 4:
MAE of item parameter estimates under MAR and NONMAR model with covariates. Polytomously scored items

Generating Values			Analysis		Mean Absolute Error				
δ	$\rho(\theta, \zeta)$	$\rho(\zeta_1, \zeta_2)$	Q_2	P	α	β_1	β_2	δ	γ
-1.0	0.0	-	-	1	0.109	0.118	0.175		
	0.4	0.8	0	1	0.139	0.133	0.212		
			1	2	0.115	0.120	0.179	0.099	0.149
			2	3	0.108	0.119	0.177	0.093	0.103
	0.8	0.8	0	1	0.143	0.159	0.254		
			1	2	0.130	0.125	0.198	0.103	0.132
			2	3	0.107	0.119	0.180	0.096	0.104
0.0	0.0	-	-	1	0.126	0.146	0.217		
	0.4	0.8	0	1	0.166	0.180	0.309		
			1	2	0.138	0.143	0.211	0.090	0.143
			2	3	0.128	0.140	0.211	0.084	0.101
	0.8	0.8	0	1	0.191	0.241	0.417		
			1	2	0.159	0.148	0.238	0.089	0.126
			2	3	0.127	0.140	0.219	0.084	0.097

mension predicted by two covariates was used, results were *MAE(α) = 0.150* and *MAE(β) = 0.116*. Further, when two dimensions with three covariates were used as a propensity model, results obtained were *MAE(α) = 0.140* and *MAE(β) = 0.114*. An analogous pattern appeared in all conditions. It can be seen that increasing the correlation of the latent variables θ and ζ, that is, increasing the violation of ignorability, resulted in more bias in the parameter estimates when the covariates are ignored. Including them reduced the bias to a value close to the baseline.

An empirical simulation

In the previous studies, all response data were simulated under IRT models. In practice, IRT models do not perfectly fit response data. For instance, the assumptions regarding local reliability, dimensionality and subgroup invariance might not be perfectly met. Therefore, an empirical simulation was carried out to investigate the effectiveness of the method with real data. The data were collected in the LISS panel of CentERdata, a representative internet panel for Longitudinal Internet Studies in the Social sciences. The panel consisted of households which are randomly selected from the municipal registers in the Netherlands. Participants filled out questionnaires on a monthly basis. The empirical simulation was based on data from two questionnaires. The first was the Mental Health Continuum-Short Form (MHC-SF, abbreviated here to MHC) (Keyes et al., 2008, Lamers et al., 2011). The MHC consisted of 14 items measuring positive mental health on a Likert scale from 0 to 5. The Dutch version of the MHC has shown good psychometric properties (Lamers et al., 2011) and the item parameters proved stable over time (Lamers et al., 2012). The second questionnaire was the Brief Symptom Inventory (BSI; Dutch version) (de Beurs & Zitman, 2006). The 53 items pertained to psychological symptoms which are scored on Likert scales ranging from 0 to 4. The sample consisted of 1,932 adults who filled out the questionnaires on four measurement occasions in nine months. The autocorrelation between response occasions was 0.653 for the MHC and 0.663 for the BSI. The distribution of item responses across the categories for both tests was quite different. The MHC had an average score of 42.55 with a maximum possible score of 70 and all response categories attracted a substantial number of responses. For the BSI the average score was low, that is, 19.30 with a maximum possible score of 212, and most responses were in the zero categories. The reason is that the BSI assesses negative psychological symptoms ranging from depression and anxiety up to paranoia and psychoticism. Such symptoms are relatively rare in the general population.

For the simulation, missing item responses were created by removing responses from the fourth occasion. This was done as follows. The probability of an observation was determined by the posterior expected estimate of θ on the previous occasion. In one condition, constant item location parameters were used to produce proportions of missing responses of 0.30 or 0.50. In a second condition, similar proportions of missing responses were created, but in this case the item location parameters were chosen in such a way that the response propensity uniformly decreased from 0.625 on the first item to 0.375 on the last item. The first condition will be labeled the Uniform condition, while the second will be labeled the Progressive condition. The means and variances of all latent dimensions were

equal to zero and one, respectively; the correlations were equal to the autocorrelations mentioned above.

100 replications were made and the item parameter estimates obtained using the complete data were the reference values to which mean absolute errors were computed using formula (9). The models used to analyze the data are listed in Table 5. All estimates were made using MML. Model 1 in Table 5 does not include a response propensity model and is expected to produce the largest bias. Model 2 has the total scores of the previous administration as covariate for θ. Model 3 has the 2PL model for the distribution of the missing data indicators but no covariates. The same held for Model 4, but here, the total

Table 5:
Summary of analysis models for empirical simulation

	Model	**Legend**
1	$p(x\mid\theta)g(\theta)$	Incomplete response data only.
2	$p(x\mid\theta)g(\theta\mid y_{t-1})$	Incomplete response data with total scores of the previous administration as covariate.
3	$p(x\mid\theta)p(d\mid\zeta)g(\theta,\zeta)$	Incomplete response data and missing data indicator.
4	$p(x\mid\theta)p(d\mid\zeta)g(\theta,\zeta\mid y_{t-1})$	Incomplete response data and missing data indicator, with total scores of the previous administration as covariate.
5	$p(x\mid\theta)p(d\mid\zeta)g(\theta,\zeta\mid\theta_{t-1})p(x_{t-1}\mid\theta_{t-1})$	Incomplete response data, missing data indicator, and responses on previous administration as covariate.

Table 6:
MAE of item parameter estimates under various MAR and NONMAR models

Measure:	**MHC**				**BSI**			
Pattern:	**Uniform**		**Progressive**		**Uniform**		**Progressive**	
Missing:	**30%**	**50%**	**30%**	**50%**	**30%**	**50%**	**30%**	**50%**
Model								
1	0.826	1.111	0.768	1.184	0.914	1.283	0.871	1.283
2	0.331	0.397	0.308	0.444	1.038	1.110	0.862	1.199
3	0.341	0.415	0.326	0.466	1.000	1.200	0.974	1.245
4	0.331	0.396	0.293	0.400	1.003	1.104	0.899	1.095
5	0.310	0.388	0.285	0.396	0.905	1.088	0.900	0.972

scores of the previous administration were a covariate for ζ. Finally, in Model 5, the latent variable pertaining to the previous administration was a covariate for ζ, while this latent variable was measured by the item response on the previous administration. Model 4 is expected to produce better results than the models 2 and 3, since this model can be seen as a combination of the two previous models. Further, Model 5 is expected to do even better. However, this expectation may be jeopardized by the increasing numbers of parameters in the models and the related increase in estimation errors.

The results are given in Table 6. For the MHC, the models 2 through 5 all created a substantial decrease in MAE from the baseline in Model 1. For instance, in the uniform condition with 30% missing data, the MAE decreased from 0.826 in Model 1 to 0.310 in Model 5. Model 3 produced the smallest decrease while Model 5 produced the highest decrease. For the MHC, this pattern was visible in all conditions. In general, the pattern was as expected: adding information decreased the effect of the violation of ignorability. Further, Model 2 was always slightly better than Model 3, so using the total score on the previous administration worked slightly better than invoking a latent variable for the pattern of missingness. Model 5 using the complete observations on the previous administration, always worked best, but overall it must be concluded that the differences between the performances of the models 2 through 4 were small.

For the BSI, the pattern was far less clear. Model 5 did produce a decrease of MAE compared to Model 1, but the decrease was small. Further, in some conditions the MAE actually increased. So here accounting for the violation of ignorability was far less successful. A tentative explanation may be related to the low average score on the BSI and the highly skewed distribution of responses across response categories with most responses in the zero categories. This seriously harms the precision of the item parameter estimates. So a conclusion here is that the effectiveness of the methods proposed here depends on the characteristics of the data.

Discussion

This article is meant as a contribution to the growing literature on modeling nonignorable missing item responses. This field of research is very broad and therefore, this article has several limitations. First of all, the focus is on the estimates of the item parameters. This topic is of interest when the item parameters are of interest, for instance in the calibration of an item bank for a computerized adaptive test or a large-scale education survey. Equally interesting are the consequences of violation of ignorability for the estimates of the person parameters, especially when these estimates have serious consequences for the tested persons. This topic deserves further study. Further, the study was limited to the generalized partial credit model, while other models such as the graded response model (Samejima, 1969) and the sequential model (Tutz, 1990) are also often used in educational and psychological measurement. The problems addressed in this article are definitely also relevant in non-parametric IRT (Sijtsma, 1998). However, it is well known that results obtained using different models and approaches to modeling polytomously scored item responses do not produce very different results (see, for instance,

Verhelst et al.), so the approach advocated here is expected to be relevant for these models also.

The next remark pertains to the implications of the presented methods for practitioners. All computations were done by MML using public domain software (MIRT, Glas, 2010b). However, this is not essential. The computations can be also done with the much-used computer program Mplus (Muthén & Muthén, 2012) or in a Bayesian framework using the computer program OpenBugs (Lunn, Spiegelhalter, Thomas & Best, 2009). Therefore, after adding missing data indicators to the data file, practitioners can perform all needed analyses within the well-known framework of latent variable and IRT modeling. The global fit of MAR and NONMAR models can be compared using the well-known statistics such as likelihood ratio statistic and its modifications, the AIC and BIC. However, also local, item oriented fit statistics dedicatedly developed for IRT models (see, for instance, Glas, 2005) can be used to evaluate the appropriateness on the models. If adding a response propensity model, possibly with covariates, leads to a substantial improvement of model fit, taking the propensity model into account is highly recommended. A final remark pertains to the fact that the simulations presented here show that applying model-based corrections to violations of ignorability can work in data-analyses with polytomously scored items, and that the propensity models can include multidimensional IRT and observed covariates. However, what the, admittedly limited, empirical study also shows, is that there is no guarantee that the approach is always effective. A tentative conclusion drawn from the empirical simulation was that unfavorable characteristics of the data, such as poor support for parameter estimates, may seriously threaten the procedure.

References

Ackerman, T.A. (1996a). Developments in multidimensional item response theory. *Applied Psychological Measurement 20,* 309-310.

Ackerman, T.A. (1996b). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement 20,* 311-329.

Adams, R.J., Wilson, M.R., & Wang, W.C. (1997). The random coefficients multinomial logit. *Applied Psychological Measurement, 21,* 1-25.

Bartholomew, D.J., & Knott, M. (1999). *Latent variable models and factor analysis* (2nd edition). London: Oxford University Press.

Bock, R.D., Gibbons, R.D., & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement, 12,* 261-280.

de Beurs, E. & Zitman, F. (2006). De Brief Symptom Inventory (BSI): de betrouwbaarheid en validiteit van een handzaam alternatief. [The Brief Symptom Inventory (BSI): the reliability and validity of a practical alternative to the SCL 90]. *Maandblad Geestelijke Volksgezondheid 61,* 120-141.

Efron, B. (1977). Discussion on maximum likelihood from incomplete data via the EM algorithm (by A. Dempster, N. Liard, & D. Rubin). *J. R. Statist. Soc.,B, 39,* 29.

Glas, C.A.W. (1992). A Rasch model with a multivariate distribution of ability. In M. Wilson (Ed.), *Objective Measurement: Theory into practice*, *Vol. 1* (pp. 236-258), New Jersey, NJ: Ablex Publishing Corporation.

Glas, C.A.W. (1999). Modification indices for the 2-pl and the nominal response model. *Psychometrika*, *64*, 273-294.

Glas, C.A.W. (2005). Assessment of model fit. In B.S.Everitt & D.C.Howel (Eds.), *Encyclopedia of Statistics in Behavioral Science*. (pp. 1243-1249). Chichester:Wiley.

Glas, C. A. W. (2010a). Item parameter estimation and item fit analysis. In W. J. van der Linden and C. A. W. Glas (Eds.), *Elements of Adaptive Testing* (pp. 269-288). New York, NJ: Springer.

Glas (2010b). MIRT, *Mulidimensional Item response Theory.* Public Domain Software. http://www.utwente.nl/gw/omd/Medewerkers/medewerkers/glas/

Glas, C.A.W., & Pimentel, J. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement, 68,* 907-922.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrika, 46,* 931-961.

Heitjan, D. F. (1994). Ignorability in general incomplete-data models. *Biometrika, 81,* 701–708.

Heitjan, D. F. (1997). Ignorability, sufficiency and ancillarity. *J. Roy. Statist. Soc. Ser. B, 59,* 375–381.

Heitjan, D. F. and Rubin, D. B. (1991). Ignorability and coarse data. *Ann. Statist. 19*, 2244–2253.

Holman,R.,& Glas, C.A.W. (2005). Modelling nonignorable missing data mechanism with item response theory models. *British Journal of Mathematical and Statistical Psychology, 58,* 1-18.

Jaeger, M. (2005). Ignorability for categorical data. *The Annals of Statistics, 33,* 1964–1981.

Keyes, C. L. M., Wissing, M., Potgieter, J. P., Temane, M., Kruger, A. & van Rooy, S. (2008). Evaluation of the mental health continuum–short form (MHC–SF) in setswana-speaking South Africans. *Clinical Psychology & Psychotherapy 15,* 181-192.

Korobko, O.B., Glas, C.A.W., Bosker, R.J. & Luyten, J.W. (2008). Comparing the difficulty of examination subjects with item response theory. *Journal of Educational Measurement, 45,* 139–157.

Lamers, S. M. A., Glas, C. A. W., Westerhof, G. J. & Bohlmeijer, E. T. (2012). Longitudinal evaluation of the Mental Health Continuum-Short Form (MHC-SF): Measurement invariance across demographics, physical illness, and mental illness. *European Journal of Psychological Assessment 28,* 290-296.

Lamers, S. M. A., Westerhof, G. J., Bohlmeijer, E. T., ten Klooster, P. M. & Keyes, C. L. M. (2011). Evaluating the psychometric properties of the mental health Continuum-Short Form (MHC-SF). *Journal of Clinical Psychology 67,* 99-110.

Little, R.J.A., & Rubin, D.B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Lord, F.M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39,* 247-264.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

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

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B, 44,* 226-233.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, *28*, 3049–3067.

Muthén, L. K., & Muthén, B. O. (2012). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Mislevy, R. J. & Wu, P. K. (1996). *Missing Responses and IRT Ability Estimation: Omits,Choice, Time limits, and Adaptive Testing.* [Research Report: RR-96-30] Educational Testing Service: Princeton (NJ).

Moustaki, I. (1996). A latent trait and a latent class model for mixed observed variables *British Journal of Mathematical and Statistical Psychology, 49,* 313-334.

Moustaki,I., & Knott, M. (2000). Weighting for item nonresponse in attitude scales by using latent variable models with covariates. *Journal of the Royal Statistical Society, A, 163,* 445-459.

Moustaki, I., & O'Muircheartaigh, C. (2000). A one dimensional latent trait model to infer attitude from nonresponse for nominal data. *Statistica,* 259-276.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16,* 159- 176.

O'Muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern Models: A latent variable approach to item Nonresponse in attitude Scales. *Journal of the Royal Statistical Society, A, 162*, 177-194.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Pohl, S., Gräfe, L. & Rose, N. (2014). Dealing With Omitted and Not-Reached Items in Competence Tests: Evaluating Approaches Accounting for Missing Responses in Item Response Theory Models. *Educational and Psychological Measurement, 74,* 423-452.

Reckase, M.D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401-412.

Reckase, M.D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W.J.van der Linden and R.K.Hambleton (Eds.), *Handbook of modern item response theory* (pp.271-286). New York, NJ: Springer.

Rose, N. (2013). *Item nonresponses in educational and psychological measurement* (Unpublished doctoral dissertation). Friedrich-Schiller-University of Jena, Germany.

Rose, N., von Davier, M., and Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT).* Research Report ETS RR-10-11, Educational Testing Service.

Rubin, D.B. (1976). Inference and missing data. *Biometrika, 63,* 581-592.

Samejima, F. (1969). Estimation of latent ability using a pattern of graded scores. *Psychometrika, Monograph Supplement, No. 17.*

Schafer,J.L., & Graham,J.W.(2002). Missing data: Our view of the state of the art. *Psychological methods, 7,* 147-177.

Sijtsma, K. (1998). Methodology Review: Nonparametric IRT Approaches to the Analysis of Dichotomous Item Scores. *Applied Psychological Measurement, 22*, 3-31.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*, 39-55.

Verhelst, N.D., Glas, C.A.W., & de Vries, H.H. (1997). A steps model to analyze partial credit. In: W.J. van der Linden and R.K. Hambleton (Eds.), *Handbook of modern item response theory*. (pp. 123-138). New York, NJ: Springer.