British Journal of Mathematical and Statistical Psychology (2005), 58, 1–17

The British Psychological Society

www.bpsjournals.co.uk

Modelling non-ignorable missing-data mechanisms with item response theory models

Rebecca Holman[1*] and Cees A. W. Glas[2]

[1]Department of Clinical Epidemiology and Biostatistics, Amsterdam Medical Center, The Netherlands

[2]Department of Educational Measurement and Data Analysis, University of Twente, Enschede, The Netherlands

A model-based procedure for assessing the extent to which missing data can be ignored and handling non-ignorable missing data is presented. The procedure is based on item response theory modelling. As an example, the approach is worked out in detail in conjunction with item response data modelled using the partial credit and generalized partial credit models. Simulation studies are carried out to assess the extent to which the bias caused by ignoring the missing-data mechanism can be reduced. Finally, the feasibility of the procedure is demonstrated using data from a study to calibrate a medical disability scale.

1. Introduction

Whenever data are collected, however carefully, the possibility, origin and treatment of missing responses should be considered. Even if care is taken to ensure that all appropriate respondents are contacted and provide some data, responses on individual variables may be missing, uncodable or in a category such as 'don't know' or 'not applicable'. If a data set contains missing observations, then the mechanism causing the incompleteness can be characterized according to its degree of randomness. Rubin (1976) described and named a number of types of mechanism.

Following Rubin, let $\tilde{\boldsymbol{d}}$ be a realization of some missing-data indicator, and let $\boldsymbol{x}_{(0)}$ and $\boldsymbol{x}_{(1)}$ be the unobserved and observed data, respectively. If the probability, $g_{\xi}(\tilde{\boldsymbol{d}}|\boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)})$, of the missing-data pattern $\tilde{\boldsymbol{d}}$ depends on neither $\boldsymbol{x}_{(0)}$ nor $\boldsymbol{x}_{(1)}$, that is if $g_{\xi}(\tilde{\boldsymbol{d}}|\boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)}) = 1$, then the data are both 'missing at random' (MAR) and 'observed at random' (OAR). Such data can also be described as 'missing completely at random' (MCAR). If $g_{\xi}(\tilde{\boldsymbol{d}}|\boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)})$ does not depend on the unobserved data $\boldsymbol{x}_{(0)}$ and the parameter of the missing-data process, $\boldsymbol{\xi}$, is distinct from the parameter $\boldsymbol{\theta}$ of the distribution of $\boldsymbol{x}_{(0)}$ and $\boldsymbol{x}_{(1)}$, then the data are MAR. If the data are MCAR or MAR, the

*Correspondence should be addressed to Dr Rebecca Holman, Department of Clinical Epidemiology and Biostatistics, Room J1B-207, Academic Medical Center, Amsterdam, The Netherlands (e-mail: r.holman@amc.uva.nl).

DOI:10.1348/000711005X47168

missing-data mechanism is ignorable for likelihood-based inferences. This means that it is not necessary to incorporate the incompleteness mechanism into models for the observed-data process. In addition, if the data are MCAR, the missing-data mechanism is ignorable for sample-based inferences. Data entry errors, lost pages of responses, and respondents following instructions incorrectly with respect to which items they should respond to are examples of mechanisms leading to MCAR data. On the other hand, in data sets where the probability of the missing-data pattern $\tilde{\boldsymbol{d}}$ depends on $\boldsymbol{\theta}$, the missing-data mechanism is not ignorable (Little & Rubin, 1987). In such cases, the missing-data mechanism must be modelled alongside the relationships of direct interest. These mechanisms can also be described in Bayesian terms. If the posterior distribution of $\boldsymbol{x}_0$ does not include a specification of the response mechanism, then the mechanism is ignorable (Rubin, 1987).

Missing data in data sets potentially suitable for analysing using item response theory (IRT) techniques can be split into four types. The first type consists of the missing observations resulting from a priori fixed incomplete test and calibration designs. Since the design is *a priori* fixed, it is inherently independent of $\boldsymbol{x}_{(0)}$ or $\boldsymbol{x}_{(1)}$, and the data are MCAR. The second type consists of a class of response-contingent designs, such as two-stage and multistage testing (Lord, 1980) and computerized adaptive testing. Here the choice of the items administered is completely governed by the responses actually observed, and independent of the unobserved responses. As a consequence, the data collected in these designs are MAR. These designs have been discussed extensively by Mislevy and Wu (1996) and Mislevy and Chang (2000).

The third and fourth types of missing data result from unscalable responses such as 'don't know' or 'not applicable' (Lord, 1974). The third type concerns situations where the scalability of the response does not depend on the latent variable to be measured. Thus the data are MAR and may also be OAR. Procedures for analysing data subject to this kind of missingness mechanism were proposed by Lord (1974), who examined the imputation of partially correct item scores, and Bock (1972), who proposed treating omitted responses as another response category. However, it has been shown that when marginal maximum likelihood estimation methods are used on data of this type, omitted responses can be ignored in the analysis (Bock & Aitkin, 1981). The fourth type of missing data are similar to the third type but result from a non-ignorable missing mechanism. This type of data may be produced when low-ability respondents fail to produce a response, as a result of discomfort or embarrassment, or simply because they have skipped items. Another example are missing responses due to time constraints. Bradlow and Thomas (1998) show that ignoring this kind of missing-data process leads to bias in parameter estimates. Therefore, the mechanism causing the incompleteness has to be included in the analysis of this type of data. Lord (1983) suggested that whether a student gave a scalable response to a particular item depended on both the ability of the examinee and a latent trait representing 'temperament'. He went on to consider ways of incorporating this information into a model. In this paper, these suggestions will be elaborated and their usefulness tested in a number of simulation studies.

This article will present four general IRT models for taking non-ignorable missing-data mechanisms into account. These models are reformulations of the models proposed by O'Muircheartaigh and Moustaki (1999, see also Moustaki & O'Muircheartaigh, 2000; Moustaki & Knott, 2000; Bernaards & Sijtsma, 1999, 2000; Conaway, 1992; Park & Brown, 1994). In the formulation presented here, the models support a simple framework to assess explicitly the extent to which the missing data are non-ingnorable. In addition, the relationship between the present models and the model

of O'Muircheartaigh and Moustaki (1999) will be outlined in more detail in an appendix on the identification of the model.

The approach presented here will be applied to the estimation of parameters in IRT models. Simulation studies will be carried out to compare the mean squared error of the estimates obtained by ignoring the missing-data process and explicitly modelling the missing-data process. Finally, the feasibility of the method will be demonstrated using data from the Amsterdam Linear Disability Score project.

2. A general IRT model for missing-data processes

Consider a two-dimensional persons by items data matrix $\boldsymbol{X}$ with entries x_{ik},$i = 1, \ldots, N$, and $k = 1, \ldots, K$. If a combination of i and k has been observed, x_{ik} is equal to the observation, otherwise it is equal to some arbitrary constant. At this point, no assumptions are made about the range of values of x_{ik}. We define a design matrix $\mathbf{D}$ of the same dimensions as X with elements

$$d_{ik} = \begin{cases} 0 & \text{if } x_{ik} \text{ was not observed,} \\ 1 & \text{if } x_{ik} \text{ was observed.} \end{cases} \tag{1}$$

Our objective is to make inferences based on a probability model $\Pi_{i,k} p(x_{ik}|d_{ik} = 1, \theta_i, \beta_k)$, where the parameters are partitioned into two sets: structural parameters β_k, $k = 1, \ldots, K$, and incidental parameters θ_i, $i = 1, \ldots, N$. Both the parameters β_k and θ_i may be vector-valued. The latter parameters are called incidental because their number increases in proportion to the number of observations, which, in general, leads to inconsistency (Neyman & Scott, 1948). It is assumed that K is a constant that does not depend on N. Kiefer and Wolfowitz (1956) have shown that, under fairly reasonable regularity conditions, consistent estimators of structural parameters can be obtained by assuming a common distribution for the incidental parameters and integrating them out of the likelihood. Inferences about the incidental parameters are then made given the estimated values of the structural parameters.

A general model for the missing-data process is given by a multidimensional IRT model (Reckase, 1985, 1997; Ackerman, 1996a, 1996b). The probability of an observation is given by

$$p(d_{ik} = 1|\xi_i \delta_k) = \frac{\exp\left(\sum_q \delta_{kq}\xi_{iq} - \delta_{k0}\right)}{1 + \exp\left(\sum_q \delta_{kq}\xi_{iq} - \delta_{k0}\right)}. \tag{2}$$

This model has the Rasch model (Rasch, 1960) and the two-parameter logistic model (2PLM: Birnbaum, 1968) as special cases. In many instances, the amount of missing data will be small, which suggests using a model with few parameters such as the Rasch model.

If the probability of a particular observation, say $p(x_{ik}|d_{ik} = 1, \theta_i, \boldsymbol{\beta})$, does not depend on $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ are independent, then the missing data are ignorable. In this situation and assuming local independence, a straightforward model for the data and the missingness processes is

$$G_1 = \prod_{i,k} p(x_{ik}|d_{ik}, \theta_i, \boldsymbol{\beta}) p(d_{ik}|\xi_i, \delta_k) g_1(\xi_i) g_2(\theta_i), \tag{3}$$

where $p(x_{ik}|d_{ik}, \theta_i, \boldsymbol{\beta})$ and $p(d_{ik}|\xi_i, \delta_k)$ are the density of the outcome variable and the design variable, respectively, and $g_1(\xi_i)$ and $g_2(\theta_i)$ are the density of ξ_i and θ_i, respectively. To model non-ignorable missing data, it will be assumed that $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ have a common distribution $g(\xi_i, \theta_i|\boldsymbol{\phi})$ that is indexed by a parameter $\boldsymbol{\phi}$, that is,

$$G_2 = \prod_{i,k} p(x_{ik}|d_{ik}, \theta_i, \boldsymbol{\beta}) p(d_{ik}|\xi_i, \delta_k) g(\xi_i, \theta_i|\boldsymbol{\phi}). \tag{4}$$

Note that $p(x_{ik}|d_{ik}, \theta_i, \boldsymbol{\beta})$ does not depend on ξ_i. The obvious alternative is that the observations may depend on $\boldsymbol{\xi}$, which leads to

$$G_3 = \prod_{i,k} p(x_{ik}|d_{ik}, \theta_i, \boldsymbol{\beta}) p(d_{ik}|\xi_i, \theta_i, \delta_k) g(\xi_i, \theta_i|\boldsymbol{\phi}), \tag{5}$$

or that the observed data depend on both $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ and the missing-data indicators on $\boldsymbol{\xi}$, that is,

$$G_4 = \prod_{i,k} p(x_{ik}|d_{ik}, \theta_i, \xi_i, \boldsymbol{\beta}) p(d_{ik}|\xi_i, \delta_k) g(\xi_i, \theta_i|\boldsymbol{\phi}). \tag{6}$$

Models (4), (5) and (6) are not necessarily different; there may be transformations of $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ that transform one model into the other. However, model (4) is conceptually the simplest: the distributions of the observed data x_{ik} and the missing-data indicator d_{ik} are parameterized by distinct sets of parameters, which have a common distribution with parameters $\boldsymbol{\phi}$. In what follows, it will be shown that the parameters $\boldsymbol{\phi}$ can be used to index the extent to which ignorability holds. The relations between the formulations (4), (5) and (6) will be outlined further below.

Using these models to analyse both the data and missingness processes together can provide extra information on the mechanisms underlying a particular data set, even if the missing-data process is ignorable. They may, for instance, give indications about the quality of the definitions of the variables collected. The models could be used in conjunction with a wide range of statistical models. However, this paper will concentrate on the use of these models in conjunction with IRT, developing the idea that the probability of a missing response depends on a separate personality trait as well as ability, proposed by Lord (1983).

3. Combination with IRT models for observed data

In the previous section, the missing-data process was modelled with an IRT model, but the model for the observations was left unspecified. In this section, it will be set out in detail how the model can be combined with an IRT model for the observed data. In that case, both θ_i and ξ_i are latent variables. The elements of matrix $\boldsymbol{D}$ will be necessarily dichotomous, whereas those in the matrix $\boldsymbol{X}$ may be either dichotomous or polytomous. Depending on the model chosen for the combined process of $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, the elements of $\boldsymbol{X}$ and $\boldsymbol{D}$ may reflect either or both latent traits.

In the example given below, both items with dichotomous and polytomous responses will be considered. These responses will be analysed with the generalized partial credit model (GPCM: Muraki, 1992). In the GPCM the probability, $p(x_{ik} = j|\theta_i, \alpha_k, \boldsymbol{\beta}_k)$, that respondent i responds to item k in category j, $j = 1, \ldots, m_k$, is denoted by

$$p(x_{ik} = j|\theta_i, \alpha_k, \boldsymbol{\beta}_k) = \frac{\exp(j\alpha_k\theta_i - \beta_{kj})}{\sum_{j=0}^{m_k}\exp(j\alpha_k\theta_i - \beta_{kj})}, \tag{7}$$

where $\boldsymbol{\beta}_k$ is a vector of item parameters $(\beta_{k0}, \beta_{k1}, \ldots, \beta_{kj}, \ldots, \beta_{km_k})$, with $\beta_{k0} = 0$ to ensure that the estimates of $\boldsymbol{\beta}_k$ are unique. The 2PLM (Birnbaum, 1968) is the special case for $m_k = 1$. Further, model (7) specializes to the partial credit model (PCM: Masters, 1982; Masters & Wright, 1997) upon setting $\alpha_k = 1$, and specializes further to the Rasch model for dichotomous items by setting $m_k = 1$. Extending (7) to include more latent traits gives

$$p(x_{ik} = j|\theta_i, \xi_i, \boldsymbol{\beta}_k) = \frac{\exp\left(j\left(\sum_q \alpha_{kq}\theta_{iq}\right) - \beta_{kj}\right)}{\sum_{j=0}^{m_k}\exp\left(j\left(\sum_q \alpha_{kq}\theta_{iq}\right) - \beta_{kj}\right)}. \tag{8}$$

Note that fixing some of the factor loadings $\delta_{k1}, \ldots, \delta_{kq}, \ldots, \delta_{kQ}$ in the model for d_{ik} given by (2) and some of the factor loadings $\alpha_{k1}, \ldots, \alpha_{kq}, \ldots, \alpha_{kS}$ in the model for x_{ik} given by (8) produces special cases of models (4), (5), and (6). For model (4) it is assumed that the responses only load on $\boldsymbol{\theta}$ and the missing-data indicators only load on $\boldsymbol{\xi}$. The factor loadings model the extent to which the missing-data indicators for item k are related to a latent overall response propensity $\boldsymbol{\xi}$. For model (5) it is assumed that the observed variables depend on $\boldsymbol{\theta}_1 = \boldsymbol{\theta}$ and the missing-data indicator variables to depend both on $\boldsymbol{\theta}_1 = \boldsymbol{\theta}$ and $\boldsymbol{\theta}_2 = \boldsymbol{\xi}$. Non-ignorability can be investigated from the loadings of the probability of missingness on $\boldsymbol{\theta}$. In model G_4 the probability of a response x_{ik} depends on an overall latent response propensity $\boldsymbol{\xi}$ which is identified because it uniquely determines the observation indicators d_{ik}. Therefore, here non-ignorability can be investigated from the loadings of the probability of the observed responses on $\boldsymbol{\xi}$. Technical details on the identification of these models are given in the Appendix.

To model the relation between the latent variables, it will be assumed that $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ have a multivariate normal distribution with density $g(\boldsymbol{\xi}, \boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The mean $\boldsymbol{\mu}$ will be set equal to zero to identify the model and the covariance matrix $\boldsymbol{\Sigma}$ is an estimand. To obtain consistent estimates, a likelihood that is marginalized with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ is maximized, that is, we maximize

$$\log L(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\Sigma}) = \sum_i \log \int \ldots \int \left[\prod_k p(x_{ik}|d_{ik}, \boldsymbol{\theta}, \boldsymbol{\beta}_k) p(d_{ik}|\boldsymbol{\xi}, \delta_k)\right] g(\boldsymbol{\xi}, \boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, \mathrm{d}\boldsymbol{\theta}\, \mathrm{d}\boldsymbol{\xi}, \tag{9}$$

with respect to the item parameters $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ and the covariance matrix $\boldsymbol{\Sigma}$. In the framework of IRT, this method is known as maximum marginal likelihood (MML; see Bock & Aitkin, 1981). The adaptation of the marginal likelihood to (3), (5) and (6) is straightforward. Procedures for maximizing this likelihood were developed by Andersen (1985), Bock, Gibbons, and Muraki (1988) and Adams, Wilson and Wang (1997). For the multidimensional Rasch and partial credit model, parameter estimates can be computed using the computer program ConQuest (Wu, Adams & Wilson, 1997); more complicated models, such as models involving (2), can be estimated using Testfact (Wilson, Wood, & Gibbons, 1991).

The model considered in (9) is closely related to a model proposed by O'Muircheartaigh and Moustaki (1999). These authors allow the observed variables to depend on the 'attitude factor', say $\boldsymbol{\theta}$ and the missing-data indicator variables to depend on two factors, say the attitude factor $\boldsymbol{\theta}$ and a factor $\boldsymbol{\xi}$. So the model is of the type given

by (5). Further, O'Muircheartaigh and Moustaki (1999) restrict the covariance matrix to an identity matrix. It is shown in the Appendix that this does not imply a restriction: the model can always be reparameterized such that the covariance matrix becomes a free estimand. An important advantage of the implication of a covariance matrix and considering models of the type given by (4) is that the correlation between $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ explicitly indexes ignorability: the more the correlation differs from zero, the more ignorability is violated. This will be used in the simulation studies presented below to assess the bias imposed when ignorability is used unjustifiably.

4. A simulation study

A simulation study was carried out to assess the effect of a missing-data process as described in equation (4) on estimates of item parameters. In this study, data were simulated using the OAR missingness process given by model (4) and analysed using the MCAR process described in model (3) and also using model (4). For sample sizes of $n = 500$, 1000, 2000, latent trait values (θ_i, ξ_i) were drawn from a bivariate normal distribution $g(\xi_i \theta_i | \Sigma)$ with means 0, variances 1 and correlation ρ, where $\rho = 0.0, 0.1, \ldots, 0.9$. These sample sizes were chosen because they reflect the numbers of respondents which could be reasonably included in a medical study similar to the AMC Linear Disability Score project used as an example in this paper. Tests consisted of $K = 10$, 20, 30 dichotomously scored items. The observations x_{ik} and the missing-data indicators d_{ik} were both generated using the Rasch model, with item parameters β_k and δ_k, respectively. The data were used to compute estimates $\hat{\beta}_k$, from model G_1, and estimates $\hat{\beta}'_k$ and $\hat{\delta}_k$ from model G_2 using MML.

The values of $\hat{\beta}_i$, $\hat{\beta}'_1$ and $\hat{\delta}_i$ were compared with the values of the parameters used to generate the data using the mean squared error (MSE) and mean absolute error (MAE). The MSE for δ_k is defined as

$$\mathrm{MSE}(\delta_k) = \sum_{r=1}^{R} \sum_{i=1}^{n} (\hat{\delta}_i - \delta_k)^2, \tag{10}$$

where $r = 1, 2 \ldots, R$ denote the replications of the simulation process and $\hat{\delta}_i$ the estimate of δ_k. The MAE is defined as

$$\mathrm{MAE}(\delta_k) = \sum_{r=1}^{R} \sum_{i=1}^{n} |\hat{\delta}_i - \delta_k|. \tag{11}$$

Ten replications were made for each combination of $K = 10$, 20, 30, $n = 500$, 1000, 2000 and $\rho = 0.1, 0.2, 0.3, \ldots, 0.9$. In the first set of simulations, the item parameters were $\delta_k = \beta_k = 0$ for all k. As a result, about 50% of the data were missing. The values of the MAE and the MSE for $\boldsymbol{\delta}$ (solid lines), $\boldsymbol{\beta}'$ (dashed lines) and $\boldsymbol{\beta}$ (dotted lines) are given in Figs. 1 and 2, respectively.

It is apparent that, for $\rho = 0.0$, the MAE and MSE of the estimates of β_k and β'_k are very close. This indicates that the estimates of β_k and β'_k are equal to within random fluctuations and confirms that the data mechanism is ignorable, and hence MCAR for $\rho = 0.0$. This is in spite of maximum likelihood estimates of IRT item parameters being slightly biased (see, for instance, Verhelst, Glas, & van der Sluis, 1984), confirming that the bias is negligible with respect to the standard errors. It can be seen that the MAE and MSE of the estimates of β_k, incorrectly assuming that the data are MCAR, increase

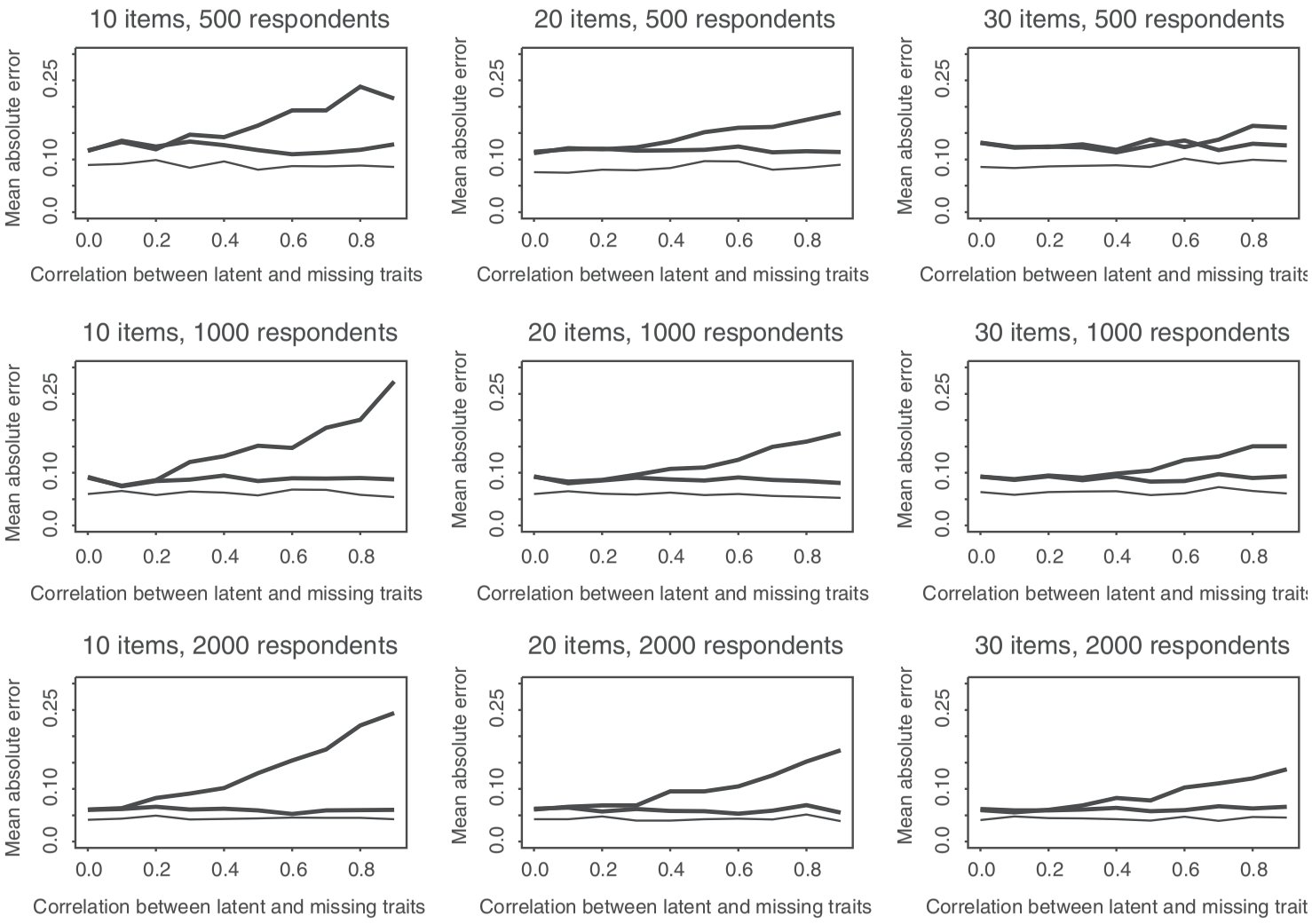

Figure 1. Comparing the values of $\hat{\boldsymbol{\delta}}$ (fine lines), $\hat{\boldsymbol{\beta}}'$ (medium lines) and $\hat{\boldsymbol{\beta}}$ (thick lines) with the values used to generate the data using the mean absolute error for $\beta_k = \delta_k = 0$.

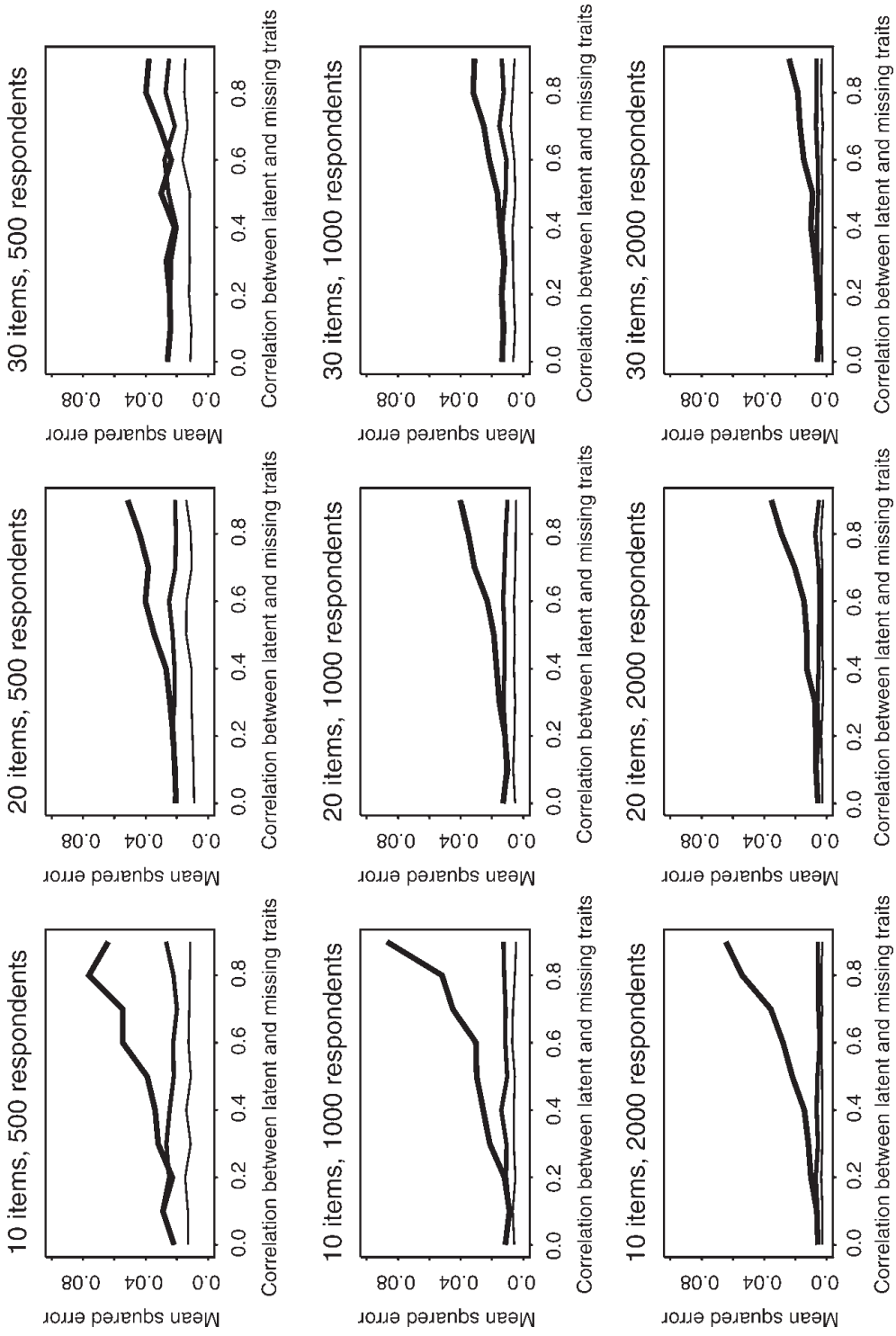

Figure 2. Comparing the values of $\hat{\boldsymbol{\delta}}$ (fine lines), $\hat{\boldsymbol{\beta}}'$ (medium lines) and $\hat{\boldsymbol{\beta}}$ (thick lines) with the values used to generate the data using the mean squared error for $\beta_k = \delta_k = 0$.

with ρ, whilst the estimates of β_k', remain stable apart from random fluctuations. This effect was most noticeable for tests with 10 items and least apparent for tests with 30 items. Partitioning the MSE into squared bias and estimation variance showed that the inflation of MSE was completely due to an inflation of bias.

In the second set of simulations the MSE and MAE were calculated over 10 replications each with $n = 1000$ respondents, $K = 30$ items, and $\delta_k = 0$ for all k for $\rho = 0.0, 0.1, \ldots, 0.9$. There sets of parameter values were chosen: for the first set $\beta_k = 1$ for all k, for the second $\beta_k = 2$ for all k and for the third $\beta_1 = \cdots = \beta_6 = 2$, $\beta_7 = \cdots = \beta_{12} = 1$, $\beta_{13} = \cdots = \beta_{18} = 0$, $\beta_{19} = \cdots = \beta_{24}^{-1}$ and $\beta_{25} = \cdots = \beta_{30} = -2$. The results of these simulations are given in Fig. 3. The relationship between ρ and the MAE and MSE of β_k are similar to those displayed in Figs. 1 and 2, indicating that the bias in the estimation of β_k caused by the non-ignorable missing mechanism is similar across different sets of item parameters.

5. The AMC Linear Disability Score project

The AMC Linear Disability Score (ALDS) project aims to develop an item bank to measure inability to perform activities of daily life (Holman, Lindeboom, Vermeulen, Glas, & de Haan, 2001; Lindeboom, Vermeulen, Holman, & de Haan, 2003). The ALDS item bank consists of about 200 items, each describing an activity that a healthy adult might perform in the course of daily life. They range from very easy (sitting up in bed) to difficult (jogging for 15 minutes). When patients are presented with the items they are asked to respond in one of three ordered response categories: 'I cannot perform the activity', 'I can perform the activity, but find it difficult' or 'I can perform the activity'. If patients have never performed an activity, then they are instructed to respond in a further category '*not applicable*'. For instance, patients who have never held a driving licence are instructed to respond to the item 'driving a car' in this way. Responses in this last category can be seen as missing, since they are not directly scalable. The data were collected by specially trained nurses.

The data used in this paper form two distinct parts of the sample being used to calibrate the item bank. The parts result from offering test 1 and test 2, each consisting of 32 items, to samples of 171 and 179 patients, respectively. The tests had no items in common. In test 1, 27 items had missing responses, and the number of missing responses per item ranged from 7 to 56, with a mean of 16.1. In test 2, 25 items had missing responses, the number ranging from 1 to 68 with a mean of 10.4.

In order to obtain an impression of the missing-data pattern, the relation between the respondents' score levels and the amount of missing responses was examined, by computing the correlation between $\text{logit}(\Sigma_k d_{ik} x_{ik} / \Sigma_k d_{ik} m_k)$ and $\text{logit}(\Sigma_k d_{ik} / K)$. Note that $m_k = 2$ for all k and $K = 32$. The correlations were -0.04 for test 1 and -0.12 for test 2. This indicates that the amount of missing responses went up with the proficiency level. A possible explanation is that patients with a higher proficiency level tended to boost their rating by failing to respond, while the patients of low proficiency were less inclined or motivated to impress the nurses.

The data were modelled using G_1, G_2, G_3 and G_4 (models (3)–(6)) and a model where the missing-data process and the observed data loaded on the same latent trait, that is, a model where the correlation between ξ_i and θ_i equals one. This model will be labelled G_0. These models were elaborated further in two versions. In the first version, the Rasch model was used to model d_{ik} and the PCM to model the observed responses x_{ik}.

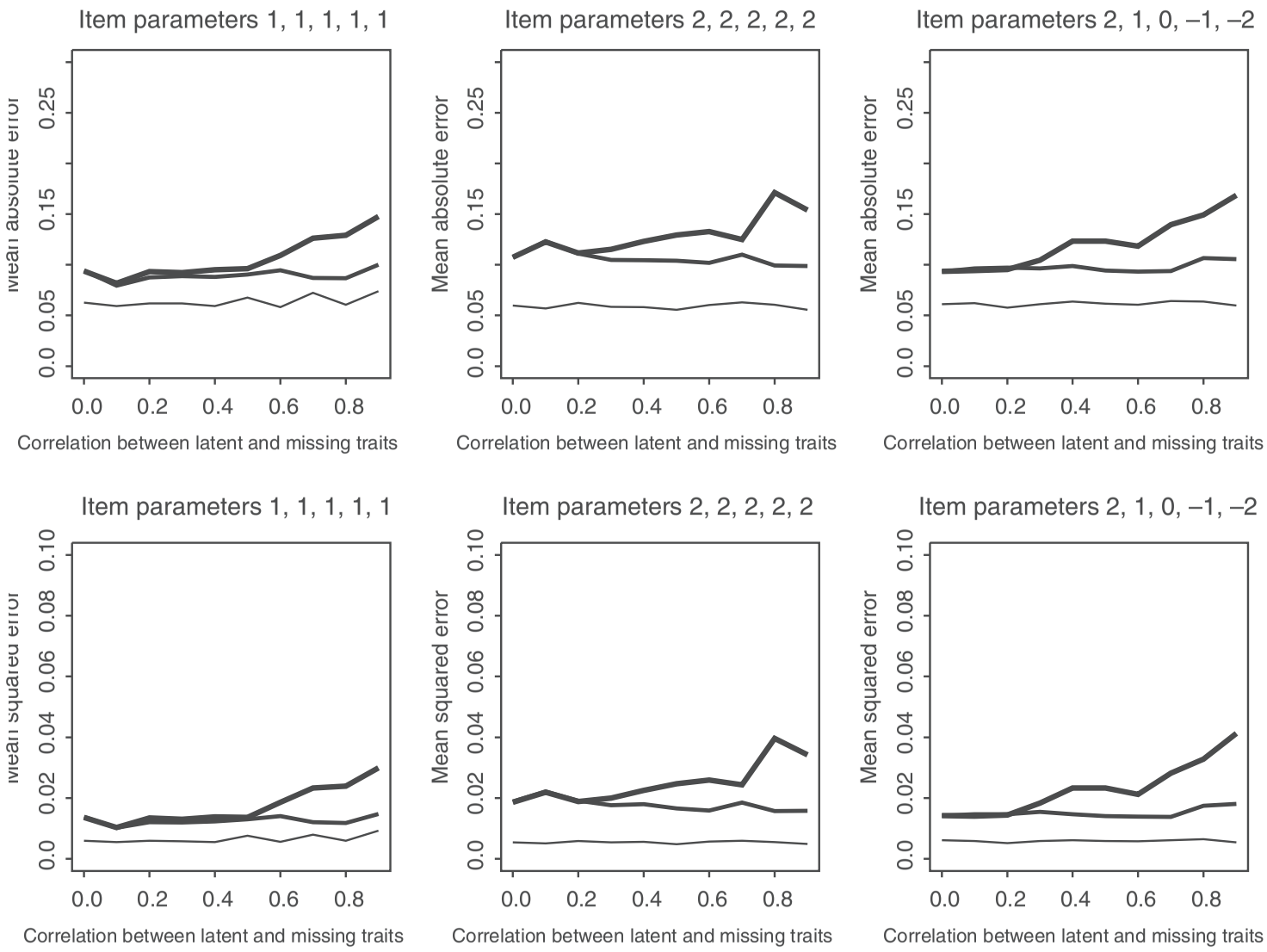

Figure 3. Comparing the values of $\hat{\boldsymbol{\delta}}$ (fine lines), $\hat{\boldsymbol{\beta}}'$ (medium lines) and $\hat{\boldsymbol{\beta}}$ (thick lines) with the values used to generate the data using the mean absolute error and mean squared error for the values of β_k given.

In this case the parameter α_i in (7) and (8) was set equal to one. In the second version, α_i was estimated, meaning that the 2PLM and the GPCM were used to model d_{ik} and x_{ik}, respectively.

An overview of the results of the analyses is given in Table 1. The last column (ρ) gives the estimates of the correlation between the dimensions for the models G_2, G_3 and G_4, and the fixed values for the other two models. Models can be compared using Akaike's information criterion (AIC) or the Bayes information criterion (BIC), which are presented in 'smaller-is-better' form. In all cases, the AIC and BIC are largest for model G_0. For the PCM with tests 1 and 2, the smallest AIC and BIC are for model G_4. For the GPCM with test 1, the smallest AIC and BIC are for G_3. However, for the GPCM with test 2, the smallest AIC is for G_4 and the smallest BIC is for G3. Finally, it can be seen that the GPCM has a better overall fit than the PCM, according to both the AIC and the BIC, but the estimates of ρ in both models are comparable. Observed correlations are usually attenuated by unreliability, which in turn is related to the number of items in the test. Comparing the estimates of the correlations in G_2 with the correlation between $\text{logit}(\Sigma_k d_{ik}x_{ik}/\Sigma_k d_{ik}m_k)$ and $\text{logit}(\Sigma_k d_{ik}/K)$ reported above, it can be seen that this is also true in the present case. In test 1 the observed correlation was -0.04, while the latent correlation was -0.124 for the PCM and -0.104 for the GPCM. In test 2 these figures were -0.12, -0.424 and -0.445, respectively. Hence, the latent correlations make the association between the proficiency level and the missing-data process more manifest.

Table 1. Results of fitting models G_0, G_1, G_2, G_3 and G_4 to the ALDS data

Data	Model		−2loglik	*k*	AIC	BIC	ρ
Test 1	PCM	G_0	11 665.7	92	11 849.7	12 138.7	1.000
		G_1	11 509.9	93	11 695.9	11 988.1	0.000
		G_2	11 508.7	94	11 696.7	11 992.0	−0.124
		G_3	11 508.7	94	11 696.7	11 992.0	−0.896
		G_4	11 495.6	94	11 683.6	11 978.9	−0.547
Test 1	GPCM	G_0	11 032.1	150	11 332.1	11 803.3	1.000
		G_1	10 950.3	150	11 250.3	11 721.5	0.000
		G_2	10 923.5	151	11 225.5	11 699.9	−0.104
		G_3	10 808.0	177	11 162.0	11 718.1	−0.678
		G_4	10 801.8	182	11 165.8	11 737.6	−0.543
Test 2	PCM	G_0	11 515.1	90	11 695.1	11 982.0	1.000
		G_1	11 308.3	91	11 490.3	11 780.4	0.000
		G_2	11 293.5	92	11 477.5	11 770.7	−0.424
		G_3	11 293.5	92	11 477.5	11 770.7	−0.758
		G_4	11 292.7	92	11 476.7	11 769.9	−0.909
Test 2	GPCM	G_0	11 295.2	146	11 587.2	12 052.6	1.000
		G_1	11 279.5	146	11 571.5	12 036.9	0.000
		G_2	11 223.5	147	11 517.5	11 986.0	−0.445
		G_3	11 168.7	171	11 510.7	12 055.7	−0.778
		G_4	11 148.5	178	11 504.5	12 071.9	−0.834

The total number of parameters estimated is denoted by *k*.
Minus twice the log-likelihood is denoted by −2loglik.

The final remark concerns model fit. Methods for the analysis of fit of multidimensional IRT models are not readily available. Therefore, two components making up the complete model, say $p(\boldsymbol{x}|\boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{\beta})$ and $p(\boldsymbol{d}|\boldsymbol{\xi}, \boldsymbol{\delta})$, were evaluated separately. The analyses were carried out using the computer program OPLM (Verhelst, Glas, & Verstralen, 1995). The program computes conditional maximum likelihood estimates of the item parameters and computes a test statistic R_{1c} (Glas & Verhelst, 1989, 1995). The test evaluates whether the item response probabilities implied by the IRT model used properly describe the observed response proportions. For this test, the score range is partitioned into a number of categories and the observed and expected response frequencies of item responses are combined into an asymptotically χ^2 distributed test statistic. If the test rejects the IRT model (say, the Rasch model or the PCM), a more general model (say the 2PLM or the GPCM) is needed. The results are shown in Table 2, in the rows labelled '1PLM' and 'PCM'.

Overall, model fit is far from perfect. The 2PLM and GPCM were then used as alternatives. Using the OPLM program the time difficulties β_{kj} were re-estimated using conditional maximum likelihood and the R_{1c} test statistic was computed as above (Glas & Verhelst, 1989, 1995). The estimation and testing procedure used in OPLM entails rounding the item discrimination parameters α_k to the nearest integer (for a motivation of this procedure, see Verhelst and Glas, 1995). Since the rounding produces an approximation to the 2PLM and the GPCM, the procedure results in a conservative test. The results are shown in Table 2, in the rows labelled 'GPCM'. It appears that the model fit for the missing-data process is now quite acceptable.

These results suggest that the missingness process in the ALDS data cannot, in general, be satisfactorily modelled using model G_0 or G_1, although models G_2 and G_3 seem to be reasonable in most cases. This suggests that the data are OAR but not MAR, meaning that the missingness process is ignorable for inferences on $\boldsymbol{\beta}$ but not for inferences on $\boldsymbol{\theta}$. In addition, the results in Table 2 suggest that the fit of the 2PLM to the missingness process is satisfactory, whereas the 1PLM is not. In addition, neither the PCM or the GPCM seems to fit the data sufficiently. This indicates that items need to be removed from the test before inferences can be made on the values of $\boldsymbol{\theta}$.

Table 2. Goodness of fit of the PCM and GPCM to the ALDS data for the latent and missingness traits in model G_1

Data	Model		Trait	R_{1c}	*df*	*p*-value
Test 1	PCM	$p(\boldsymbol{x}\|\boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{\beta})$	latent	317.8	189	0.0000
	1PLM	$p(\boldsymbol{d}\|\boldsymbol{\xi},\boldsymbol{\delta})$	missingness	112.2	78	0.0063
	GPCM	$p(\boldsymbol{x}\|\boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{\beta})$	latent	248.9	189	0.0021
	2PLM	$p(\boldsymbol{d}\|\boldsymbol{\xi},\boldsymbol{\delta})$	missingness	60.6	69	0.7694
Test 2	PCM	$p(\boldsymbol{x}\|\boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{\beta})$	latent	399.4	189	0.0000
	1PLM	$p(\boldsymbol{d}\|\boldsymbol{\xi},\boldsymbol{\delta})$	missingness	63.4	48	0.0657
	GPCM	$p(\boldsymbol{x}\|\boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{\beta})$	latent	392.6	189	0.0000
	2PLM	$p(\boldsymbol{d}\|\boldsymbol{\xi},\boldsymbol{\delta})$	missingness	20.4	22	0.5782

6. Discussion and conclusion

A variety of methods for dealing with ignorable and non-ignorable missing data in practical situations have been proposed (Schafer, 1997). These range from imputation

methods to algorithms which permit parameters to be estimated, whilst ignoring missing observations. The development of models in which the primary data and missingness processes are considered jointly (Heckman, 1979) is particularly interesting. These models can be useful in situations where it is thought that the mechanism causing the missing data is not ignorable. A model-based procedure for handling non-ignorable missing data using IRT models is presented that is formulated in such a way that the extent to which ignorability is violated can be easily assessed. Four general IRT models for missing-data mechanisms are proposed. As an example, these models are worked out in detail in conjunction with item response data modelled by the partial credit and generalized partial credit models. In a number of simulation studies it was shown that ignoring the missing-data process results in considerable bias in the estimates of the item parameters. This bias increases as a function of the correlation between the proficiency to be measured and the latent variable governing the missing-data process. Further, it was shown that this bias can be reduced using the models presented above. The feasibility of the procedure was demonstrated using data from a calibration study of a medical disability scale. The correlation between the proficiency and the latent variable of the missing-data processes was significant, and using the missing-data models significantly increased model fit.

This approach can be generalized by the inclusion of covariates in the missing data model. IRT models with manifest covariates were proposed by Zwinderman (1991, 1997) and Adams *et al.* (1997). Finally, test statistics are needed to evaluate the appropriateness of the models presented above. This provides another incentive for the development of evaluation methods for fit to multidimensional IRT models.

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 model. *Applied Psychological Measurement*, *21*, 1–25.

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*, 3–16.

Bernaards, C. A., & Sijtsma, K. (1999). Factor analysis of multidimensional polytomous item response data suffering from ignorable item nonresponse. *Multivariate Behavioral Research*, *34*, 277–313.

Bernaards, C. A., & Sijtsma, K. (2000). Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research*, *35*, 321–364.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*, *46*, 443–459.

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

Bradlow, E. T., & Thomas, N. (1998). Item response theory models applied to data allowing examinee choice. *Journal of Educational and Behavioral Statistics*, *23*, 236–243.

Conaway, M. R. (1992). The analysis of repeated categorical measurements subject to nonignorable nonresponse. *Journal of the American Statistical Association*, *87*, 817–824.

Glas, C. A. W., & Verhelst, N. D. (1989). Extensions of the partial credit model. *Psychometrika*, *54*, 635–659.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds), *Rasch models: foundations, recent developments and applications*. New York: Springer-Verlag.

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

Holman, R., Lindeboom, R., Vermeulen, M., Glas, C. A. W., & de Haan, R. J. (2001). The Amsterdam Linear Disability Score (ALDS) Project. The calibration of an item bank to measure functional status using item response theory. *Quality of Life Newsletter*, *27*, 3–4.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, *27*, 887–903.

Lindeboom, R., Vermeulen, M., Holman, R., & de Haan, R. J. (2003). Activities of daily living instruments in clinical neurology: optimizing scales for neurologic assessments. *Neurology*, *60*, 738–742.

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. (1983). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika*, *48*, 477–482.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds), *Handbook of modern item response theory*. New York: Springer-Verlag.

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin*, *115*, 300–307.

McDonald, R. P. (1982). Linear versus nonlinear models in item reponse theory. *Applied Psychological Measurement*, *6*, 379–396.

Mislevy, R. J., & Chang, H. H. (2000). Does adaptive testing violate local independence? *Psycometrika*, *65*, 149–156.

Mislevy, R. J., & Wu, P. K. (1996). *Missing responses and IRT ability estimation: omits, choice, time limits, adaptive testing*. ETS Research Report RR-96-30-ONR, Princeton, NJ: Educational Testing Service.

Moustaki, I., & Knott, M. (2000). Weighting for item non-response in attitude scales by using latent variable models with covariates. *Journal of the Royal Statistical Society, Series 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*, *60*, 259–276.

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

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*, 1–32.

O'Muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: a latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society, Series A*, *162*, 177–194.

Park, T., & Brown, M. B. (1994). Models for categorical data with nonignorable nonresponse. *Journal of the American Statistical Association*, *89*, 44–52.

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

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 & R. K. Hambleton (Eds), *Handbook of modern item response theory*. New York: Springer-Verlag.

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

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman & Hall.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408.

Verhelst, N. D., & Glas, C. A. W. (1995). The generalized one parameter model: OPLM. In G. H. Fischer & I. W. Molenaar (Eds), *Rasch models: their foundations, recent developments and applications*. New York: Springer-Verlag.

Verhelst, N. D., Glas, C. A. W., & van der Sluis, A. (1984). Estimation problems in the Rasch model: the basic symmetric functions. *Computational Statistics Quarterly*, *1*, 245–262.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). *OPLM: computer program and manual*. Arnhem, The Netherlands: Cito.

Wilson, D. T., Wood, R., & Gibbons, R. (1991). *TESTFACT: Test scoring, item statistics, and item factor analysis (computer software)*. Chicago: Scientific Software International, Inc.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1997). *ConQuest: Generalized item response modelling software*. Australian Council for Educational Research.

Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika*, *56*, 589–600.

Zwinderman, A. H. (1997). Response models with manifest predictors. In W. J. van der Linden & R. K. Hambleton (Eds), *Handbook of modern item response theory*. New York: Springer-Verlag.

Received 5 March 2002; revised version received 11 February 2004

Appendix: Identification and relations between various formulations of the model

A multidimensional IRT model will be considered for variables $y_{ik} (i = 1, \ldots, N, k = 1, \ldots, K)$, where y_{ik} assumes values $(0, 1, \ldots, M)$ and

$$p(y_{ik} = j) = \frac{\exp\left(j \sum_{q=1}^{Q} \alpha_{kq} \theta_{iq} - \beta_{kj}\right)}{1 + \sum_{h=1}^{M} \exp\left(h\left(\sum_{q=1}^{Q} \alpha_{kq} \theta_{iq}\right) - \beta_{kh}\right)}. \tag{A1}$$

The variables y_{ik} can be both the observations x_{ik} and the missingness indicators d_{ik}. The latent variables $\boldsymbol{\theta}_i, \boldsymbol{\theta}_i' = (\theta_{i1}, \ldots, \theta_{iq}, \ldots, \theta_{iQ})$, have a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. McDonald (1982; see also Mellenbergh, 1994; Takane & de Leeuw, 1987) points out that this model can be viewed as a factor analysis model with Q factors, where $\theta_{i1}, \ldots, \theta_{iq}, \ldots, \theta_{iQ}$ are factor scores and $\alpha_{11}, \ldots, \alpha_{kq}, \ldots, \alpha_{KQ}$ are factor loadings.

The model can be identified in two ways:

1. using the restrictions $\boldsymbol{\mu} = \mathbf{0}$, and $\alpha_{jq} = 1$, if $j = q$, and $\alpha_{jq} = 0$, if $j \neq q$, for $j = 1, \ldots, Q$ and $q = 1, \ldots, Q$, in which case $\boldsymbol{\Sigma}$ is a free estimand; or
2. using the restrictions $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Sigma} = \mathbf{I}$, and $\alpha^{o}_{jp} = 0$, for $j = 1, \ldots, Q - 1$ and $q = j + 1, \ldots, Q$.

Let $\mathbf{A}$ and $\mathbf{A}^{o}$ be the matrices of discrimination parameters for the former and latter parametrization, respectively. That is, $\mathbf{A}$ is defined as a $K \times Q$ matrix with elements α_{iq}, and $\mathbf{A}^{o}$ is defined analogously. For example with $K = 5$ and $Q = 3$, the first parametrization results in a matrix $\mathbf{A}$ given by

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \alpha_{41} & \alpha_{42} & \alpha_{43} \\ \alpha_{51} & \alpha_{52} & \alpha_{53} \end{bmatrix}, \tag{A2}$$

and a free matrix Σ, while the second parameterization results in a matrix $\boldsymbol{A}$ given by

$$\mathbf{A}^{o} = \begin{bmatrix} \alpha^{o}_{11} & 0 & 0 \\ \alpha^{o}_{21} & \alpha^{o}_{22} & 0 \\ \alpha^{o}_{31} & \alpha^{o}_{32} & \alpha^{o}_{33} \\ \alpha^{o}_{41} & \alpha^{o}_{42} & \alpha^{o}_{43} \\ \alpha^{o}_{51} & \alpha^{o}_{52} & \alpha^{o}_{53} \end{bmatrix}, \tag{A3}$$

and a covariance matrix Σ that is equal to the identity matrix. In both cases, the number of restrictions is equal to Q^2. In the first example $\mathbf{A}$ has nine restrictions, while in the second, $\mathbf{A}^{o}$ has three restrictions and Σ has three restrictions on diagonal and three on off-diagonal elements.

The parameters $\boldsymbol{\theta}_i$ can be transformed to $\boldsymbol{\theta}^{o}_i$ by $\boldsymbol{\theta}^{o}_i = \mathbf{L}^{-1}\boldsymbol{\theta}_i$, where $\mathbf{L}$ comes from the Cholesky decomposition of Σ. Because $\mathbf{L}$ is lower triangular and $\mathbf{A}\boldsymbol{\theta}_i = \mathbf{A}\mathbf{L}\boldsymbol{\theta}^{o}_i = \mathbf{A}^{o}\boldsymbol{\theta}^{o}_i$, the restrictions $\alpha_{jq} = 1$, if $j = q$, and $\alpha_{jq} = 0$, if $j \neq q$, for $j = 1, \ldots, Q$ and $q = 1, \ldots, Q$, are transformed into restrictions $\alpha^{o}_{jq} = 0$, for $j = 1, \ldots, Q-1$ and $q = j+1, \ldots, Q$. On the other hand, defining the lower triangular matrix $\mathbf{F}$ as the first Q rows of $\mathbf{A}^{o}$ and applying $\boldsymbol{\theta}_i = \mathbf{F}\boldsymbol{\theta}^{o}_i$, results in $\Sigma = \mathbf{F}\mathbf{F}^{T}$ and $\mathbf{A} = \mathbf{A}^{o}\mathbf{F}^{-1}$, which in turn produces restrictions $\alpha_{jq} = 1$, if $j = q$, and $\alpha_{jq} = 0$, if $j \neq q$, for $j = 1, \ldots, Q$ and $q = 1, \ldots, Q$. Hence, the two parameterizations of the model are easily interchanged.

As already mentioned, in the application discussed above, the variables y_{ik} can either be the observations x_{ik} or missing-data indicators d_{ik}. For instance, suppose that the Rasch model holds for both the observations x_{ik} and missing data indicators d_{ik}. Then, for a test with three items, the matrix $\mathbf{A}$ can be defined as

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{A4}$$

where rows 1, 3 and 5 relate to responses and the rows 2, 4, and 6, to missing-data indicators. The responses load on the first factor only, and the missing-data indicators load on the second factor only. The correlation between the factors indexes the dependence between the observations and the missing-data indicators, that is, the extent to which the missing-data process is ignorable.

In the formulation given by O'Muircheartaigh and Moustaki (1999), the covariance matrix is an identity matrix. In that case, the model in this example can be identified using

$$\mathbf{A} = \begin{bmatrix} \alpha_{11} & 0 \\ \alpha_{21} & \alpha_{22} \\ \alpha_{31} & 0 \\ \alpha_{41} & \alpha_{42} \\ \alpha_{51} & 0 \\ \alpha_{61} & \alpha_{62} \end{bmatrix}. \tag{A5}$$

Note that both the responses and the missing-data indicators load on both dimensions. The model can be transformed to a version with a free covariance matrix, but this would, in most cases, not lead to separate dimensions for the responses and missing-data indicators. Hence, models with this property are a special case of the general model proposed by O'Muircheartaigh and Moustaki (1999).