Psychological Test and Assessment Modeling, Volume 57, 2015 (4), 595-618

Bayesian estimation in IRT models with missing values in background variables

Christian Aßmann[1], *Christoph Gaasch*[2], *Steffi Pohl*[3] & *Claus H. Carstensen*[2]

Abstract

Large scale assessment studies typically aim at investigating the relationship between persons competencies and explaining variables. Individual competencies are often estimated by explicitly including explaining background variables into corresponding Item Response Theory models. Since missing values in background variables inevitably occur, strategies to handle the uncertainty related to missing values in parameter estimation are required. We propose to adapt a Bayesian estimation strategy based on Markov Chain Monte Carlo techniques. Sampling from the posterior distribution of parameters is thereby enriched by sampling from the full conditional distribution of the missing values. We consider non-parametric as well as parametric approximations for the full conditional distributions of missing values, thus allowing for a flexible incorporation of metric as well as categorical background variables. We evaluate the validity of our approach with respect to statistical accuracy by a simulation study controlling the missing values generating mechanism. We show that the proposed Bayesian strategy allows for effective comparison of nested model specifications via gauging highest posterior density intervals of all involved model parameters. An illustration of the suggested approach uses data from the National Educational Panel Study on mathematical competencies of fifth grade students.

Keywords: missing values, background variables, classification and regression trees, item response theory

[1] *Correspondence concerning this article should be addressed to:* Dr. Christian Aßmann, Otto-Friedrich-University Bamberg, Feldkirchenstr. 21, 96045 Bamberg, Germany; email: christian.assmann@uni-bamberg.de

[2] Leibniz Institute for Educational Trajectories Bamberg, Germany; Otto-Friedrich-Universität Bamberg, Germany

[3] Free University Berlin, Germany

1 Introduction

With large scale assessments, such as the Program for International Student Assessment (PISA; e.g., OECD, 2012), the Third International Mathematics and Science Study (TIMSS; e.g., Mullis, Martin, Foy, & Arora, 2012), the National Assessment of Educational Progress in the United States (NAEP; e.g., National Center for Education Statistics, 2013) or the German National Educational Panel Study (NEPS; e.g., Blossfeld, Roßnach, & Maurice, 2011), researchers aim at investigating the relationship between competencies and explaining variables. Typical research questions concern, for example, the explanation of competencies and competence development based on individual characteristics like gender, socio-economic status, migration background, and context variables like school characteristics. Competencies in large scale studies are assessed via tests (see e.g. OECD, 2012; Weinert, et al., 2011) and competence data are usually analyzed via Item Response Theory (IRT) models. IRT models belong to the group of Confirmatory Item Factor Analysis (CIFA) models (see Edwards, 2010). These models include, for example, the Rasch model (Rasch, 1960), the Partial Credit model (Masters, 1982), the 3-parameter logistic model (Birnbaum, 1968), or the corresponding model counterparts based on the normal ogive description (see, e.g., Albert, 1992; Béguin & Glas, 2001). All these modeling approaches have in common that parameter estimation within these frameworks routinely relies on the Likelihood principle either in form of Maximum Likelihood (ML) estimation typically performed via the EM-algorithm, or in form of Bayesian approaches using Markov Chain Monte Carlo (MCMC) based techniques (see Edwards, 2010). Since the derived likelihood functions or moment equations involve high-dimensional integrals or incorporate latent structures, the considered model frameworks are in principle straightforward to handle within a Bayesian estimation approach using MCMC techniques. Further, all these different modeling frameworks offer the possibility to be extended to incorporate auxiliary information in form of context (background) variables of the examinees into the estimation of ability parameters (Mislevy, 1987). Incorporating auxiliary information in form of context variables enhances efficiency of estimates and allows for direct assessment of dependencies between the latent abilities and context variables.

Despite tremendous efforts in field work, missing values in these background variables occur. Usually these missing values are treated via multiple imputation (Rubin, 1987) including relevant variables in the imputation model to capture dependencies. Specifically questionnaire variables (i.e., background variables) are needed for the provision of competence scores (in form of plausible values) and questionnaire variables as well as latent competence scores are needed for the imputation of missing responses in questionnaire items. Other large scale studies, such as PISA and NAEP, deal with this problem by using missing indicators for each questionnaire variable (indicating whether the response on this variable is observed or missing) and aggregating the questionnaire variables and the response indicators to orthogonal factors. The set of factors is then used as background variables in the IRT measurement model of the competence data (see Allen, Carlson, Johnson & Mislevy, 2001). Thereby, as many factors as needed to explain 90 percent of the variance of the questionnaire items are typically considered. However, this

approach is a two-step approach that does not incorporate the latent competence score in the imputation of the questionnaire variables and does not depict the uncertainty stemming from missing values in questionnaire items.

We propose to extend available Bayesian estimation routines relying on MCMC methodology and augment the parameter vector with the missing values in the background variables. The advantages of the suggested approach compared to a two-step approach relate to increased statistical efficiency and model consistency. Whilst the current model presupposes the distribution of latent abilities to depend on background variables, this dependence could not be reflected to the full extent by a two-step approach using some pre estimated latent ability for the imputation of background variables. Fully accounting for the assumed conditional variables further increases the statistical accuracy in terms of efficiency in assessing the influence of background variables competencies. As the background variables are not subject to specific modelling, flexible ad hoc assumptions are required to provide a valid approximation of the underlying full conditional distribution of the missing values. While parametric normal models, as used in Aßmann, Gaasch, Pohl, & Carstensen (2016), offer flexible handling of missing values in metric background variables, background variables in large-scale assessments are typically also categorically scaled. Hence, in this paper we suggest to adapt non-parametric approximations to the full conditional distributions based on sequential regression trees. As discussed within the literature (Burgette & Reiter, 2010; Doove, van Buuren, & Dusseldorp, 2014), these are also able to model more complex dependencies, for example, higher order interactions. The MCMC approach allows furthermore for direct assessment of accuracy measures of estimators without requirement to use combining rules typically needed when analyzing data sets with missing values.

In the following we will first describe the IRT framework and the Bayesian estimation routines used in our approach. We will then present our hybrid sampling scheme and demonstrate its performance first in a simulation study investigating the estimation accuracy and second by an empirical example demonstrating the applicability of the approach.

2 Item Response Theory for scaling of competence tests

In large scale assessments often different competence domains are assessed, for example reading, mathematical competence, science, and computing literacy. The competence domains are assessed by tests that contain a number of items that may be dichotomously or polytomously scored. Different types of IRT models are used in the different large scale assessment studies. While PISA (OECD, 2012) and NEPS (Pohl & Carstensen, 2012) rely on one parameter IRT models such as the Rasch model (Rasch, 1960) or the partial credit model (Masters, 1982), NAEP (National Center for Education Statistics, 2013) and TIMSS (Mullis, Martin, Foy, & Arora, 2012) use two or three parameter IRT models (Birnbaum, 1968). The multidimensional random coefficients multinomial logit model is a general model which many large scale assessment studies rely on (e.g., PISA [OECD, 2012] and NEPS [Pohl & Carstensen, 2012]). To illustrate the suggested estima-

tion strategy based on the device of data augmentation, we consider a simplified version of the multidimensional random coefficients multinomial logit model. In order to reduce the computational burden for estimation, we refer to binary responses only, and use the probit link to model the individual response probability. The considered model states the probability for person i to answer item j correctly as

$$\Pr\left(y_{ij}=1|\theta_i\right)=\Phi\left(\alpha_j\theta_i-\xi_j\right) \quad i=1,...,N \quad j=1,...,J\,, \tag{1}$$

where $\Phi(\bullet)$ denotes the standard normal cumulative distribution function, Y_{ij} refers to the response given by person i on item j, θ_i to the ability of person i, α_j, $j=1,\ldots,J$ denote the discrimination parameter, and ξ_j, $j=1,\ldots,J$ the item difficulty parameter. For completion, to solve the inherent non-identifiability of the parameters, the sum of the item difficulties is set to equal zero, that is $\sum_{j=1}^{J}\xi_j=0$ and for the discrimination parameters $\prod_{j=1}^{J}\alpha_j=1$ with $\alpha_j>0$ for all $j=1,\ldots,J$. Note, that missing values may also occur within the competence test items. These are usually scored as wrong response, partially correct, ignored in the estimation, or the missing process is explicitly modelled (for a discussion of the different approaches see Pohl et al. (2014)). The suggested approach to deal with missing values in background variables is general and applicable to any chosen approach on dealing with missing competence items. Further, θ_i is regarded as a random parameter with density function $g(\theta_i)$ for all i. Commonly, the population distribution $g(\theta_i)$ is assumed normal with mean μ and variance σ^2. Assuming a mixing distribution for θ_i allows for handling the identification problem arising in case of treating θ_i as a fixed individual specific parameter. This model allows to simultaneously model item responses and structural relations by allowing the inclusion of explaining variables for the latent competence variable. If such explaining variables (background variables) are included in the model, the distribution $g(\bullet)$ is assumed normal with mean $Z_i\gamma$, where Z_i denotes a vector of Q individual characteristics (background variables) influencing individual ability. This corresponds to the multivariate regression equation

$$\theta_i=Z_i\gamma+\epsilon_i, \text{ with } \epsilon_i\sim N\left(0,\sigma^2\right). \tag{2}$$

Substituting θ_i according to this regression setup into Equation (1) results in $\Phi\left(\alpha_j\theta_i-\xi_j\right)=\Phi\left(\alpha_jZ_i\gamma+\alpha_j\epsilon_i-\xi_j\right)$. The statistical analysis of this model framework is non-trivial, when missing values occur in the background variables.

3 The proposed approach for dealing with missing values in person background variables within IRT models

In the following we will describe a data analysis strategy that applies to univariate competence measurement settings with missing values in background variables. Thus, the proposed approach is applicable to different IRT models including the Rasch model and the two parameter logistic or normal model. The approach is presented for binary response variables. The proposed estimation routine is designed to cope with missing information on individual level variables influencing person abilities. These background variables may be metric or categorical. We adopt a Bayesian estimation scheme that allows for a conceptually stringent treatment of missing values in observed individual characteristics via the device of data augmentation (see Tanner & Wong, 1987). Bayesian estimation is implemented using MCMC techniques, namely Gibbs sampling, which are ideally suited to deal with the hierarchical structure of the model and the handling of missing values. In addition, the usage of MCMC simulation methods proves straightforward for complex IRT models relative to marginal maximum likelihood as discussed in (Patz & Junker, 1999). Summarizing all parameters as ψ for a given model and letting S denote the sample data, Bayesian inference is concerned with the posterior distribution $p(\psi \mid S)$ and corresponding moments thereof. Gibbs sampling is a device to produce a sample from the joint posterior distribution of the parameter vector ψ, which can be used to estimate posterior moments and density estimates. Posterior draws of ψ partitioned into convenient blocks $\psi = \{\psi_t\}_{t=1}^{T}$ are obtained via Gibbs sampling, when direct sampling from the posterior distribution is difficult, but sampling from the full conditional distributions is directly accessible. The functional forms of the full conditional distributions can be deduced from the joint posterior distribution of ψ and S, that is $p(\psi, S) = L(S \mid \psi)\pi(\psi)$, where $L(S \mid \psi)$ denotes the model likelihood and $\pi(\psi)$ denotes the a priori distribution, via isolating the kernel of a single parameter block ψ_t conditional on all other blocks $\psi_1, \ldots, \psi_{t-1}, \psi_{t+1}, \ldots, \psi_T$ and the data S. Since the functional forms of the full conditional distributions depend on the assumed prior distributions, these are in general conveniently chosen to facilitate sampling from closed form full conditional distributions. Among others Liu, Taylor, G., & Belin (2000) and Schafer (1999) propose Gibbs sampling as a special MCMC technique to incorporate uncertainty of missing values into parameter estimation. This tool has been applied among others by Koskin, Robins, & Pattison (2010) in the context of Bayesian estimation of social network models. Missing values are thereby incorporated within the parameter vector as a parameter block of their own. The corresponding MCMC scheme providing a sample from the posterior distribution can be adapted to include also the full conditional distributions of the missing values. While several possibilities exist to specify a full conditional distribution for missing values in the explaining factors, a parsimonious yet flexible alternative to parametric models is offered via non-parametric sequential regressions as suggested by Burgette & Reiter (2010). Note that next to non-parametric approaches,

also semi-parametric approaches based on chained equations are available (see van Buuren & Groothuis-Oudshoorn, 1987).

3.1 Estimation algorithm

To illustrate the proposed treatment of missing values, we use the IRT model framework for binary items outlined in Equations (1) to (2), which allows for closed form sampling from the full conditional distributions employed within the Gibbs sampler along the lines suggested by Albert (1992) and Edwards (2010).[4] For illustrative purposes the model takes the following form. As stated beforehand, missing values may also occur in competence items. In order to provide the likelihood function of observed item responses, we define a missing indicator t_{ij} taking value one if a response of individual i to item j is missing and zero otherwise. Summarizing all parameters as ψ, all binary responses as Y, and all background variables as Z, then the corresponding likelihood is given as

$$\mathcal{L}(Y|\psi,Z)=\prod_{i=1}^{N}\int\left[\prod_{j=1}^{J}(\Phi\Big(\big(2y_{ij}-1\big)\big(\alpha_j\theta_i-\xi_j\big)\Big)\big(1-t_{ij}\big)+t_{ij}\right]g(\theta_i\,|\,Z_i)d\theta_i \qquad (3)$$

where the mixing distribution $g(\theta_i\,|\,Z_i)$ relates to the regression setup given in Equation (2). Given this model setup, the corresponding set of full conditional distributions can be described as follows. The derivation of the full conditional distributions follows the mechanistic principles outlined by Albert (1992) and also Chib (2001). Note that in addition to augmenting the parameter vector by the auxiliary variable y_{ij}^{*} arising from the latent linear regression providing the probability rationale as stated in Equation (4) for observed responses, that is

$$y_{ij}=\begin{cases}1 \text{if } y_{ij}^{*}>0,\\ 0 \text{else},\end{cases} \text{ where } y_{ij}^{*}=\alpha_j\theta_i-\xi_j+e_{ij}, \quad \text{with} \quad e_{ij}\overset{iid}{\sim}N\big(0,1\big), \qquad (4)$$

the parameter vector is also augmented with the missing values in background variables. As the assumed model is not concerned with the background variables, the likelihood from Equation (3) is not informative with regard to the full conditional distribution of missing values in the background variables. Instead, we suggest either a parametric or a non-parametric ad-hoc approximation to the full conditional distribution of missing values, which is then added to the set of full conditional distributions. This allows the researcher to account for the uncertainty created by the missing values directly in the background variables. Considering the typically occurring categorical context variables (such as gender, school type, or migration background), such non-parametric approximations seem especially suited to provide valid characterizations of the underlying full conditional distributions. As pointed out by Burgette & Reiter (2010) as well as Doove,

[4] See also (Aßmann & Boysen-Hogrefe, 2011) for a general treatment of Bayesian estimation for binary panel probit models.

van Buuren, & Dusseldorp (2014), the flexibility of non-parametric approaches to cope also with non-linear dependency and non-standard interactions possibly the challenges inherent to statistical analysis with missing data.

After initializing parameters as draws from the prior distributions obeying the identifying restrictions, the following iterative scheme with repetitions $r = 1,\ldots,R$ arises for repetition *r*.

Step 1

The underlying latent variable y_{ij}^* is sampled from a truncated normal distribution with corresponding parameters

$$\mu_{y_{ij}^*} = \alpha_j \theta_i - \xi_j, \quad \text{and} \quad \sigma_{y_{ij}^*} = 1,$$

where truncation sphere is $(-\infty, 0)$ for $y_{ij} = 0$ and $(0, \infty)$ for $y_{ij} = 1$. Note that step 1 is performed for all observed answers, that is for all *i* and *j* with $t_{ij} = 0$.

Step 2

For sampling discrimination and item parameters, we summarize them as $\varphi_j = (\alpha_j, \xi_j)$ as suggested by the equation

$$y_j^* = X_j \varphi_j + e_j,$$

where $y_j^* = T_j \left(y_{1j}, \ldots, y_{Nj}^* \right)'$ and $X_j = T_j (\theta - 1)$, with T_j denoting a $\sum_{i=1}^{N} (1 - t_{ij}) \times N$ matrix selecting from all individuals those with observed responses to item *j*, and θ denotes the stacking of individual competencies into a $N \times 1$ vector.[5] This allows for sampling from a bivariate normal distribution with mean vector and covariance matrix given as

$$\mu_{\varphi j} = (X_j' X_j + \Omega_{\varphi j}^{-1})^{-1} \left(X_j' y_j^* + \Omega_{\varphi j}^{-1} m_{\varphi j} \right), \text{ and } \sum_{\varphi j} (X_j' X_j + \Omega_{\varphi j}^{-1})^{-1},$$

where $m_{\varphi j}$ and $\Omega_{\varphi j}$ are the mean and variance of the conjugate bivariate normal prior distribution. To take the identifying restrictions into account, the drawn parameters denoted as $\tilde{\alpha}_j$ and $\tilde{\xi}_j$ are restricted after each iteration of the Gibbs sampler to provide

---

[5] Consider as an example. $y_j^* = (-3.5\ NA\ NA\ 3.5)'$ .. Then with

$$T_j = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

we have $T_j y_j^* = (-3.5\ 3.5)'$.

new draws used in the successive steps, that is $\alpha_j = \frac{\tilde{\alpha}_j}{(\prod_j^J \tilde{\alpha}_j)^{1/J}}$ and $\xi_j = \tilde{\xi}_j - \frac{1}{J}\sum_{j=1}^{J}\tilde{\xi}_j$.

Sampling is done for all $j = 1,\ldots J$.

Step 3

The individual abilities θ_i are sampled from a full conditional normal distribution. With T_i denoting a $\sum_{j=1}^{J}\left(1-t_{ij}\right)\times J$ matrix selecting all items with observed responses from individual *i*, define $\tilde{y}_i = T_i\left(y_i^* - \xi\right)$ and $\tilde{X}_i = T_i\alpha$, where α, y_i^* and ξ denote the vectors of all *J* discrimination parameters, latent variables or difficulties. This allows for stating the moments and variance of the full conditional multivariate normal distribution as

$$\mu_{\theta_i} = (\widetilde{X'}_i\tilde{X}_i + \sigma^{-2})^{-1}\left(\widetilde{X'}_i\tilde{y}_i + \sigma^{-2}Z_i\gamma\right) \text{and } \sigma^2_{\theta_i} = (\widetilde{X'}_i\tilde{X}_i + \sigma^{-2})^{-1}.$$

Sampling is performed for all $i = 1,\ldots,N$. In accordance with the assumed model structure, individual abilities θ_i are sampled conditional on observed responses and background variables.

Step 4

Draws from full conditional distribution for γ are obtained from a multivariate normal distribution with corresponding moments given as

$$\mu_\gamma = \left(\frac{Z'Z}{\sigma^2} + \Omega_\gamma^{-1}\right)^{-1}\left(\frac{Z'\theta}{\sigma^2} + \Omega_\gamma^{-1}\nu_\gamma\right) \text{ and } \Sigma_\gamma = \left(\frac{Z'Z}{\sigma^2} + \Omega_\gamma^{-1}\right)^{-1}$$

with *Z* denoting the matrix of background variables, ν_γ and covariance matrix Ω_γ denoting the moments of corresponding conjugate normal prior distribution.

Step 5

Choosing the independent conjugate prior for σ^2 as inverse gamma with parameters a_0 and b_0, then the full conditional of σ^2 is also distributed inverse gamma with corresponding parameters

$$\tilde{a} = \frac{N}{2} + a_0, \text{ and } \quad \tilde{b} = \left(\frac{1}{2}\sum_{i=1}^{N}\left(\theta_i - Z_i\gamma\right)^2 + b_0\right)^{-1}.$$

Step 6

As the parameter vector is augmented with the missing values in background variables, corresponding draws are either obtained based on parametric approximations to the full conditional distributions or on non-parametric approximations to the full conditional

distributions. While a parametric approximation is efficient for dealing with metric background variables and known dependencies, the non-parametric approach allows for flexible handling of categorical background variables and is capable to account for non-linear dependencies among the variables. The following set of conditioning variables denoted as matrix $W_q = \left(\iota, Z_{-q}, \theta, SC\right)$ is considered, where ι denotes a vector of N ones, Z_{-q} the matrix of background variables except variable Z_q, and *SC* denotes a vector of sample statistics, for example the sum score, of the competence items for each individual. Note that the draws for the missing values in the background variables are based upon the competence score θ and, thus, the latent ability is incorporated in the imputation. Note that if no missing values are present within the responses to the competence items, using the sum score as a sample statistic of observed item responses may not enrich the model with further information, as the sum score may be subsumed within the vector of latent abilities θ.

a) When using a parametric approximation to the full conditional distribution of the missing values, draws for the missing values in the $N \times Q$ matrix of background variables *Z* can be obtained via specifying a univariate normal full conditional distribution for each of the *Q* variables contained in *Z*. Within the intercourse of the Gibbs sampler, imputed and hence complete variables are at hand for each iteration *r*, resulting in the following *Q* regression equations given as

$$Z_q = W_q \varphi_q + \epsilon_q,\ q = 1,\ldots,Q\ .$$

Imputations are then generated as follows. Each missing value in Z_q is replaced via a draw from a univariate normal distribution with expectation $W'_{mis}\hat{\varphi}$ and variance $\widehat{\sigma_\epsilon^2}$. To account for the uncertainty of the least squares estimators $\hat{\varphi}$ and $\widehat{\sigma_\epsilon^2}$, draws from the corresponding asymptotic distributions are used for generating draws for the missing values in Z_q. Further, it should be explicitly noted that the estimation scheme introduces the updated draws of the individual latent abilities θ into the imputation model for each iteration and that a parametric scheme becomes computationally burdensome for missing values in discrete variables with many categories.

b) In order to establish flexible approximations to the full conditional distributions of missing values in background variables *Z*, Burgette & Reiter (2010) suggest to use nonparametric approximations obtained via Classification and Regression Trees CART (see Breiman, Friedman, Olshen, & Stone, 1984). The flexibility of the CART algorithm to incorporate nonlinear dependencies among the variables with missing data has been highlighted by Doove et al. (2014). CART constitutes a non-parametric recursive partition algorithm. The objective of CART is to split up the observations into different groups, fulfilling the condition that respondents and thus observations assigned to one group show highest intra-group homogeneity with respect to the relevant variable, whereas the inter-group homogeneity is intended to be as small as possible. There exist manifold possibilities to define a partition. CART

defines binary partitions via a set of conditioning variables. In the present application the set of conditioning variables containing all available information is given via W_q. To ensure computational feasibility, the CART algorithm does consider univariate splits only, that is only binary partitions defined upon a single variable are considered.
The sequential partitioning algorithm proceeds via consideration of all binary partitions. Of all possible binary partitions defined by univariate splits, the split with the maximum reduction in heterogeneity is selected as the optimal partition. As indicators for heterogeneity the variance in case of metric and the entropy in case of categorical variables are chosen. The resulting binary partition of the data along the set of conditioning variables provides sets of admissible values defining the nonparametric characterization of the full conditional distribution and serving as donors for filling in the missing values. All respondents can be assigned to one of these identified donor groups. Each missing value is imputed via a draw from the empirical distribution within this donor group using a Bayesian bootstrap. Thus, the uncertainty of the unobserved missing values is directly taken into account in parameter estimation. With regard to the settings of the CART algorithm, concerning stopping criteria and minimum requirements for the size of donor groups, we follow the suggestions of Burgette & Reiter (2010). Hence, no further split is considered when the resulting reference groups contain less than 50 or the gain in homogeneity is less than 0.01.

Given a sample of all model parameters obtained via iterative sequential cycling through the set of full conditional distributions, the plausible values for each individual can be directly taken from the provided Gibbs output. After discarding an appropriate burn-in phase each sweep $\{\theta_i\}_{r=1}^{R}$, $i=1,\ldots,n$ from the posterior distribution could be taken as a vector of plausible values. The proposed approach simultaneously deals with the estimation of plausible values and the imputation of missing values in background variables, thus, accounting for both sources of uncertainty.

3.2 Simulation study

To assess the statistical validity of the proposed approach, we set up a simulation design investigating the statistical accuracy of parameter estimation. Within this design we compare the Bayesian data augmentation approach to handle missing values with the Bayesian full sample estimates before deletion and those from a Bayesian complete case analysis in which cases with missing values in background variables are deleted listwise. Further, we assess the parametric and the non-parametric approximation towards the full conditional distribution of missing values. The before deletion situation is thereby chosen as a benchmark providing estimators with highest possible statistical accuracy for a given data set, whilst the complete case situation illustrates reduced efficiency in parameter estimation when ignoring missing values. The simulation study builds upon

repeated estimation of simulated data sets. Different data and missing values generating mechanisms are considered. The data sets with missing values in the background variables are generated based on the complete data sets. Then for each of the complete data sets the corresponding Bayesian estimates before deletion are calculated, while for each of the data sets with missing values the Bayesian estimates based on data augmentation for handling of missing values are calculated. The estimates using data augmentation are compared to the estimates before deletion and the complete cases, when all individuals with missing values in the background variables are removed from the data set.

The detailed conditions of the data generating and missing values generating processes are as follows. For each of the $C = 200$ replications, the binary response pattern to competence items is simulated using the model stated in Equations (1) and (2) with a sample setup of $N = 1000$ individuals facing $J = 10$ items. The item difficulty parameters are fixed across the replications. To fulfil the implemented identifying restriction $0 = \sum_{j=1}^{J} \xi_J$ they are derived as $\xi_j = \tilde{\xi}_j - \frac{1}{J}\sum_{j=1}^{J}\tilde{\xi}_j$, where $\tilde{\xi}_j \sim N(0, 0.5),\ 1, \ldots J$ are draws from a normal distribution. Correspondingly, to ensure the positivity and scaling constraint on the discrimination parameters, these are generated as transformed draws from a lognormal distribution, that is $\alpha_j = \frac{u_j}{\left(\prod_{j=1}^{J} u_j\right)^{1/J}}$ with $u_j \sim LN(3, 0.25)$ for $j = 1, \ldots, J$. For each data set, four background variables $Z_q, q = 1, \ldots, 4$, explaining differences in individual abilities θ_i are generated from a multivariate normal distribution, where each variable has mean of zero, unit variance, and pairwise correlation equal to 0.5. The following transformations were applied to the variables. Z_2 is squared, Z_3 is transformed into a dichotomous variable with split at a value of 0, and Z_4 is categorized into a 4-way categorical variable by its quartiles. The regression coefficients including an intercept are set to $\gamma = (1.0, 0.5, -0.5, 0.5, -0.25, -0.5, -1)$ for the corresponding set of background variables given as

$$Z = \left(1, Z_1, Z_2^2, I_{\{0,1\}}(Z_3 < 0), I_{\{0,1\}}(Z_4 < q_{25}), I_{\{0,1\}}(q_{25} < Z_4 < q_{50}), I_{\{0,1\}}(q_{50} < Z_4 < q_{75})\right) .$$

The individual abilities are then generated as draws from a normal distribution with mean as implied by $Z\gamma$ and a conditional variance parameter of $\sigma^2 = 1.44$. Given values for the individual abilities, the latent variables $y_{ij}^* = \alpha_j \theta_i - \xi_j + e_{ij}$ are calculated with e_{ij} given as independent draws from a standard normal distribution and then dichotomized according to $y_{ij} = I_{\{0,1\}}(y_{ij}^* < 0)$.

Missing values are generated as follows. We distinguish two missing data generating mechanisms, which differ in their severity. With regard to data generating mechanism I

on average 5% missing values in Z_1 and 10% missing values in Z_2 were generated completely at random. For data generating mechanism II the rates of missingness increase to 10% and 20% and depend on variable Z_3 according to $\Pr(Z_{i1} = NA) = \Phi(-0.5 - 0.5Z_{i3})$, and $\Pr(Z_{i2} = NA) = \Pr(Z_{i4} = NA) = \Phi(-1.5 + 0.5Z_{i3})$. Thus, data generating mechanism II poses more challenges on the estimation routine than data generating mechanism I. Further note, that the data generating mechanism I is in line with the parametric approximation of the full conditional distribution for the augmented missing values, while missing data generating mechanism II with missing values in the categorical variable is not accessible to the parametric approximation of the full conditional distribution of the missing values.

Each of the repeated estimations for the total of $C = 200$ data sets is based on MCMC sequences with each having a length of 2000 iterations. Inspection of time series and autocorrelation plots for all parameters indicates convergence. After discarding the first quarter of the samples as burn-in, inference is based on the remaining 1500 simulated draws from the joint posterior distribution. For evaluation of the results, the parameter estimates using the data augmentation approach are compared to those before deletion and complete cases via inspection of average estimates, the root mean squared errors, and the proportion of 95% highest posterior density regions that contain the true parameter values, that is coverages.

For the missing data generating mechanism I Table 1 shows the true parameter values, the means of the posterior expected values, their standard deviations, root mean squared errors (RMSE), and coverages over $C = 200$ replications for all considered estimation approaches, that is the before deletion (BD), the complete case analysis (CC), the parametric (PI), and the non-parametric (NPI) imputation method. For the BD estimators we find overall unbiased estimation results for all parameters with all corresponding coverages reaching their nominal level as expected. Also, average standard deviations and RMSE coincide, thus highlighting the statistical accuracy of estimation in the case without missing values. Similar results are revealed for the PI estimators. Note that the implemented parametric imputation model using full conditional normal distributions reflects completely the chosen simulation setup making use of the normal distribution. Hence, in this setup the PI procedure is the most efficient way to handle missing values in the estimation. In this sense, the statistical accuracy of the PI approach serves as a benchmark for the NPI approach. Within the considered simulation study, the NPI approach reveals unbiased estimation of all parameters. Further, inspection of the statistical accuracy in terms of the RMSE and coverages suggest no severe loss of statistical accuracy compared to the BD and the parametric approach. For example, the mean standard deviation of the regression parameter γ_4 almost equals its RMSE, which is in line with the findings for the BD and PI estimators. Also with respect to the coverage rates, the findings support that there is no notable difference to the BD estimators reported in the first block columns of Table 1. The observed number of intervals covering the particular parameter corresponds to their expected theoretical values, for example

Table 1:
Missing data mechanism I - MCAR

		mean				sd				RMSE				coverage			
Parameter	true	BD	CC	PI	NPI	BD	CC	PI	NPI	BD	CC	PI	NPI	BD	CC	PI	NPI
γ1	1.000	1.010	1.022	1.013	0.991	0.108	0.129	0.111	0.111	0.107	0.122	0.112	0.110	0.945	0.960	0.950	0.960
γ2	0.500	0.501	0.505	0.500	0.494	0.052	0.062	0.055	0.055	0.050	0.060	0.052	0.051	0.965	0.960	0.970	0.975
γ3	-0.500	-0.506	-0.510	-0.501	-0.496	0.037	0.043	0.038	0.039	0.038	0.043	0.040	0.040	0.955	0.950	0.945	0.950
γ4	0.500	0.502	0.507	0.504	0.494	0.097	0.115	0.099	0.099	0.095	0.114	0.099	0.096	0.945	0.940	0.940	0.940
γ5	-0.250	-0.248	-0.263	-0.251	-0.233	0.126	0.151	0.129	0.129	0.122	0.147	0.129	0.128	0.950	0.955	0.965	0.945
γ6	-0.500	-0.506	-0.518	-0.509	-0.487	0.131	0.156	0.134	0.134	0.141	0.157	0.144	0.143	0.935	0.945	0.925	0.935
γ7	-1.000	-0.992	-1.003	-0.995	-0.971	0.143	0.170	0.146	0.147	0.151	0.172	0.157	0.156	0.935	0.940	0.910	0.920
σ^2	1.440	1.465	1.489	1.454	1.476	0.104	0.125	0.107	0.107	0.102	0.133	0.100	0.105	0.960	0.945	0.975	0.970
α1	1.295	1.302	1.311	1.303	1.303	0.098	0.117	0.098	0.098	0.104	0.124	0.104	0.105	0.930	0.950	0.935	0.910
α2	1.218	1.235	1.235	1.244	1.236	0.102	0.120	0.102	0.102	0.112	0.134	0.113	0.115	0.930	0.930	0.930	0.925
α3	0.709	0.712	0.713	0.715	0.712	0.052	0.062	0.053	0.052	0.056	0.062	0.056	0.056	0.950	0.955	0.955	0.945
α4	1.007	1.012	1.009	1.009	1.012	0.070	0.083	0.070	0.070	0.071	0.085	0.070	0.071	0.950	0.960	0.950	0.950
α5	1.043	1.041	1.044	1.039	1.042	0.072	0.085	0.072	0.072	0.075	0.085	0.073	0.074	0.950	0.960	0.950	0.945
α6	1.229	1.243	1.246	1.243	1.243	0.092	0.110	0.092	0.093	0.097	0.121	0.098	0.098	0.945	0.925	0.930	0.935
α7	1.186	1.203	1.207	1.196	1.202	0.091	0.109	0.092	0.091	0.097	0.114	0.099	0.098	0.925	0.935	0.920	0.920
α8	0.879	0.878	0.883	0.876	0.878	0.060	0.072	0.060	0.060	0.065	0.076	0.065	0.066	0.930	0.945	0.935	0.940
α9	0.909	0.918	0.923	0.921	0.918	0.066	0.079	0.066	0.066	0.072	0.083	0.072	0.072	0.920	0.950	0.930	0.930
α10	0.731	0.726	0.725	0.726	0.726	0.049	0.058	0.049	0.049	0.050	0.061	0.050	0.051	0.950	0.910	0.955	0.940
ξ1	-0.177	-0.181	-0.183	-0.181	-0.181	0.056	0.066	0.056	0.056	0.051	0.064	0.051	0.051	0.965	0.935	0.965	0.960
ξ2	-1.010	-1.018	-1.028	-1.021	-1.019	0.075	0.088	0.075	0.074	0.076	0.098	0.075	0.077	0.950	0.930	0.965	0.935
ξ3	-0.751	-0.757	-0.758	-0.755	-0.757	0.052	0.062	0.052	0.052	0.054	0.065	0.055	0.055	0.955	0.960	0.945	0.950
ξ4	0.630	0.636	0.636	0.636	0.637	0.061	0.073	0.061	0.061	0.061	0.073	0.060	0.061	0.965	0.955	0.960	0.955
ξ5	0.471	0.468	0.474	0.469	0.469	0.058	0.069	0.058	0.058	0.058	0.070	0.058	0.059	0.940	0.940	0.940	0.940
ξ6	-0.312	-0.325	-0.329	-0.325	-0.325	0.056	0.066	0.056	0.056	0.063	0.072	0.064	0.063	0.930	0.945	0.930	0.940
ξ7	1.021	1.044	1.054	1.040	1.043	0.081	0.096	0.080	0.080	0.090	0.112	0.090	0.090	0.915	0.910	0.910	0.915
ξ8	0.622	0.621	0.624	0.621	0.621	0.058	0.069	0.058	0.058	0.061	0.070	0.061	0.061	0.945	0.950	0.940	0.940
ξ9	-0.711	-0.709	-0.712	-0.708	-0.709	0.056	0.066	0.056	0.056	0.052	0.059	0.052	0.051	0.980	0.970	0.980	0.975
ξ10	0.219	0.222	0.223	0.223	0.222	0.049	0.057	0.049	0.048	0.048	0.056	0.049	0.048	0.955	0.960	0.960	0.965

Notes: Average posterior means and standard deviation, root mean squared error and coverage of C=200 replications for a data set before deletion (BD), complete case analysis (CC), the parametric imputation method (PI), and the non-parametric imputation method (NPI).

Table 2:
Missing data mechanism II - MAR

		mean			sd			RMSE			coverage		
Parameter	true	BD	CC	NPI	BD	CC	NPI	BD	CC	NPI	BD	CC	NPI
γ1	1.000	1.010	1.011	0.964	0.108	0.146	0.116	0.107	0.146	0.122	0.945	0.935	0.930
γ2	0.500	0.501	0.506	0.485	0.052	0.073	0.056	0.050	0.072	0.058	0.965	0.950	0.935
γ3	−0.500	−0.506	−0.512	−0.494	0.037	0.052	0.040	0.038	0.051	0.040	0.955	0.945	0.960
γ4	0.500	0.502	0.503	0.488	0.097	0.137	0.100	0.095	0.125	0.098	0.945	0.955	0.955
γ5	−0.250	−0.248	−0.230	−0.204	0.126	0.172	0.140	0.122	0.182	0.150	0.950	0.930	0.915
γ6	−0.500	−0.506	−0.492	−0.453	0.131	0.181	0.146	0.141	0.193	0.167	0.935	0.925	0.925
γ7	−1.000	−0.992	−0.995	−0.931	0.143	0.200	0.161	0.151	0.200	0.178	0.935	0.965	0.895
σ^2	1.440	1.465	1.500	1.482	0.104	0.149	0.109	0.102	0.163	0.111	0.960	0.930	0.950
α1	1.295	1.302	1.311	1.303	0.098	0.139	0.097	0.104	0.146	0.105	0.930	0.955	0.920
α2	1.218	1.235	1.243	1.236	0.102	0.142	0.100	0.112	0.163	0.111	0.930	0.930	0.930
α3	0.709	0.712	0.713	0.713	0.052	0.073	0.052	0.056	0.077	0.056	0.950	0.945	0.940
α4	1.007	1.012	1.019	1.012	0.070	0.099	0.071	0.071	0.097	0.071	0.950	0.955	0.970
α5	1.043	1.041	1.051	1.041	0.072	0.101	0.072	0.075	0.105	0.075	0.950	0.950	0.935
α6	1.229	1.243	1.249	1.243	0.092	0.129	0.092	0.097	0.135	0.099	0.945	0.940	0.930
α7	1.186	1.203	1.203	1.201	0.091	0.129	0.091	0.097	0.141	0.098	0.925	0.925	0.930
α8	0.879	0.878	0.884	0.878	0.060	0.085	0.060	0.065	0.091	0.065	0.930	0.920	0.950
α9	0.909	0.918	0.928	0.918	0.066	0.093	0.066	0.072	0.107	0.072	0.920	0.915	0.930
α10	0.731	0.726	0.727	0.726	0.049	0.068	0.049	0.050	0.066	0.050	0.950	0.975	0.945
ξ1	−0.177	−0.181	−0.185	−0.181	0.056	0.077	0.055	0.051	0.073	0.051	0.965	0.955	0.965
ξ2	−1.010	−1.018	−1.030	−1.018	0.075	0.104	0.074	0.076	0.108	0.075	0.950	0.940	0.930
ξ3	−0.751	−0.757	−0.760	−0.756	0.052	0.073	0.052	0.054	0.070	0.054	0.955	0.960	0.950
ξ4	0.630	0.636	0.647	0.637	0.061	0.085	0.061	0.061	0.091	0.060	0.965	0.940	0.955
ξ5	0.471	0.468	0.476	0.469	0.058	0.080	0.058	0.058	0.082	0.058	0.940	0.940	0.940
ξ6	−0.312	−0.325	−0.323	−0.326	0.056	0.077	0.056	0.063	0.087	0.063	0.930	0.910	0.940
ξ7	1.021	1.044	1.050	1.042	0.081	0.112	0.080	0.090	0.122	0.090	0.915	0.930	0.910
ξ8	0.622	0.621	0.629	0.621	0.058	0.080	0.058	0.061	0.081	0.061	0.945	0.960	0.950
ξ9	−0.711	−0.709	−0.722	−0.709	0.056	0.078	0.056	0.052	0.077	0.051	0.980	0.960	0.980
ξ10	0.219	0.222	0.219	0.222	0.049	0.067	0.048	0.048	0.063	0.048	0.955	0.945	0.960

Notes: Average posterior means and standard deviations, root mean squared error and coverage of C=200 replications for a data set before deletion (BD), complete case analysis (CC), and the non-parametric imputation method (NPI).

a 95% binomial proportion interval allows the coverages to lie between [0.907; 0.993] at $C = 200$ replications for all considered parameters. For all considered estimation approaches, the 95% highest density intervals concur with this range.

The results for the simulation study considering the missing data to occur at random are presented in Table 2. Similarly to the results from the missing completely at random simulation design, the NPI method performs very well to estimate item parameters and the structural relations on individual abilities. Overall, inspection of the simulation results suggests again high statistical accuracy of the suggested data augmentation approach. For instance, the mean standard deviations of the regression parameters differ from their corresponding RMSEs only modestly. This is supported by the reported coverage rate, which meets the expected level of 95% for all estimated parameters.

Given the evidence from the simulation study, we conclude that the nonparametric approach provides a valid and highly flexible approximation to the full conditional distribution of missing values. Given the almost negligible loss in statistical accuracy compared to the benchmark BD and PI estimators, we suggest that the use of data augmentation using non-parametric devices is a suitable solution for dealing with partially observed background variables in the context of IRT models, regardless of the scaling type of background variables.

3.3 Empirical application

To illustrate the applicability of the suggested estimation approach, we provide an exemplary empirical analysis, where we use data from the NEPS cohort sample of students in fifth grade (Blossfeld, Roßnach, & Maurice, 2011). For a description of the assessment of mathematical competence in NEPS, see Weinert, et al. (2011) and Neumann, Durchhardt, Ehmke, Grüßing, & Knopp (2013). Duchhardt & Gerdes (2012) provides an overview of the respective competence data and Skopek, Pink, & Bela (2013) the data manual of the corresponding scientific use file. The data used in this analysis contains student information on the first panel wave in the year 2010. We restrict our sample to the cases where parental information is available and to students with a valid response to at least one of the $J = 23$ binary mathematics test items. The sample considered in our analyses consists of $\mathrm{N} = 3615$ students and their parents. Consistent with the scaling in the NEPS (Pohl & Carstensen, 2012; Pohl, Gräf, & Rose, 2014), missing values on test items set are ignored. To check the robustness of this approach, we further provide the estimation results based on cases with complete background variables only (N = 2955).

We consider several background variables on student level to gauge their impact on mathematical competence. These are gender, a dummy variable indicating if the student is a German native speaker or not, students mathematical self-concept, and satisfaction with school. Additionally, we include years of education of the parent who is responsible for the everyday issues of the child and the amount of books at home (ordinal variables) as indicators of the socio-economic and cultural context. Descriptive statistics for the data considered in the application are displayed in Table 3. With at most 6% univariate

Table 3:
Descriptive statistics for backround variables.

Variable	min	max	mean	sd	missing
categorical					
female	0	1	0.48	-	0%
non German native speaker	0	1	0.11	-	2%
number of books	0	5	-	-	5%
continuous					
self-concept	1	4	2.95	0.84	6%
school satisfaction	0	10	7.74	2.94	3%
parent years of education	8	18	13.90	2.94	4%

Notes: N=3615

missing values, namely for mathematical self-concept, the total amount of missing data is considered to be relatively small.

We apply the proposed data augmented Gibbs sampling approach based on the nonparametric approximation of the full conditional distribution of missing values for estimating the regression coefficients relating background variables and the latent mathematics competence. The trace plots show no indication of convergence problems (Figure 1, Figure 2, and Figure 3). The observed autocorrelations are low, and also the cumulative means indicate no converge problems at all. Taken a burn-in period of 2000 draws, all parameter estimates are based on 8000 simulated draws. Table 4 depicts the estimated posterior means and standard deviations, as well as the 95% highest density intervals. While the results indicate a lower level of competence for females and non German native speakers, mathematical self-concept, and school satisfaction have a positive effect on students mathematical abilities. Both, a higher level of parental education and a higher number of books available in the household, are positively associated with students' mathematical ability (with all other predictors assumed constant). Note that the regression coefficients reflect the relationship of questionnaire variables with latent mathematics scores that are purified from measurement error. The estimated standard errors of the regression coefficients incorporate not only the uncertainty due to person sampling, but also uncertainty due to missing values in the predictors. Comparison with the results obtained for cases with complete background variables reveals no substantial differences. However, all parameter estimates based on complete cases only have higher standard deviations. This illustrates the increased efficiency of the suggested data augmentation approach to handle missing values in background variables.

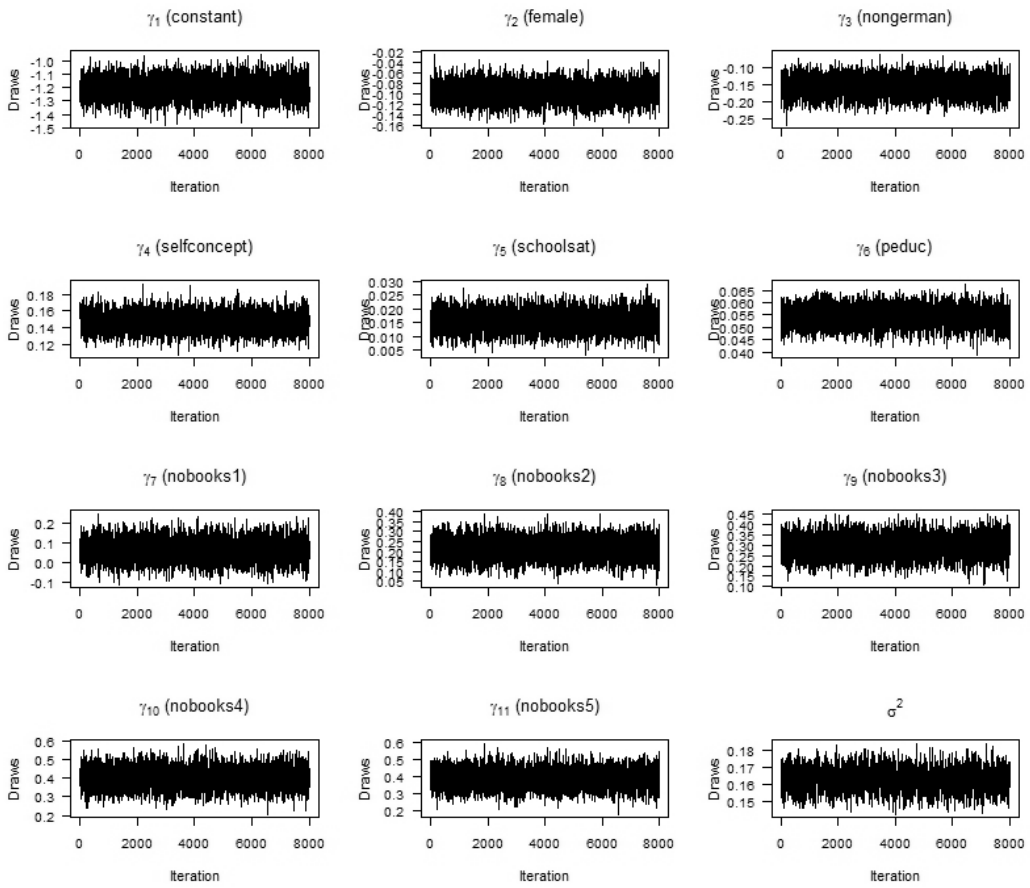

Figure 1:
Trace plots for the regression constant (γ_1), the regression coefficients for sex (γ_2), German native speaker (γ_3), mathematical self-concept (γ_4), school satisfaction (γ_5), years of education parent (γ_6), number of books at home ($\gamma_7, \gamma_8, \gamma_9, \gamma_{10}, \gamma_{11}$), as well as the residual variance (σ^2).

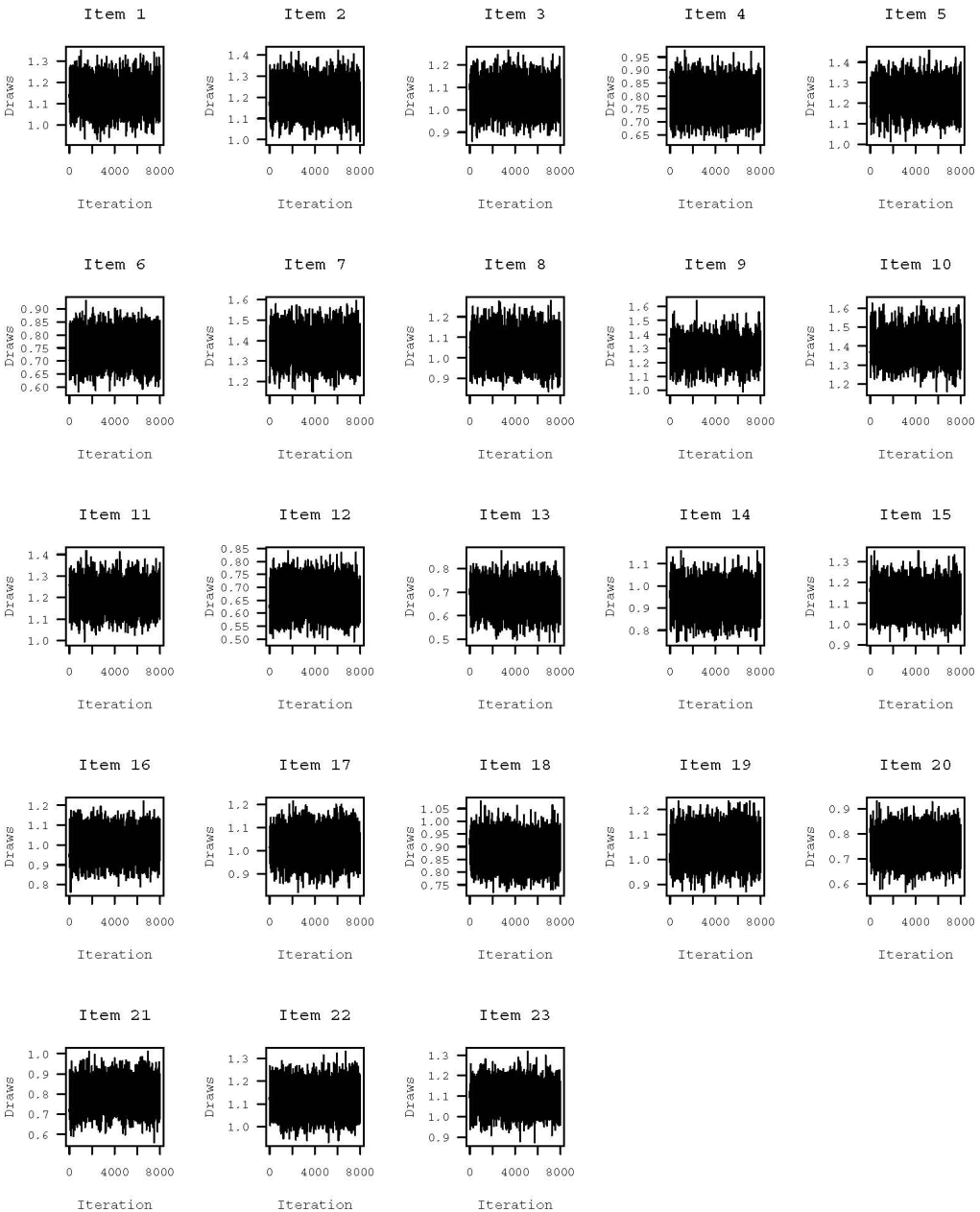

Figure 2:
Trace plots for the discrimination parameters (α_j, $j = 1,\ldots,J$).

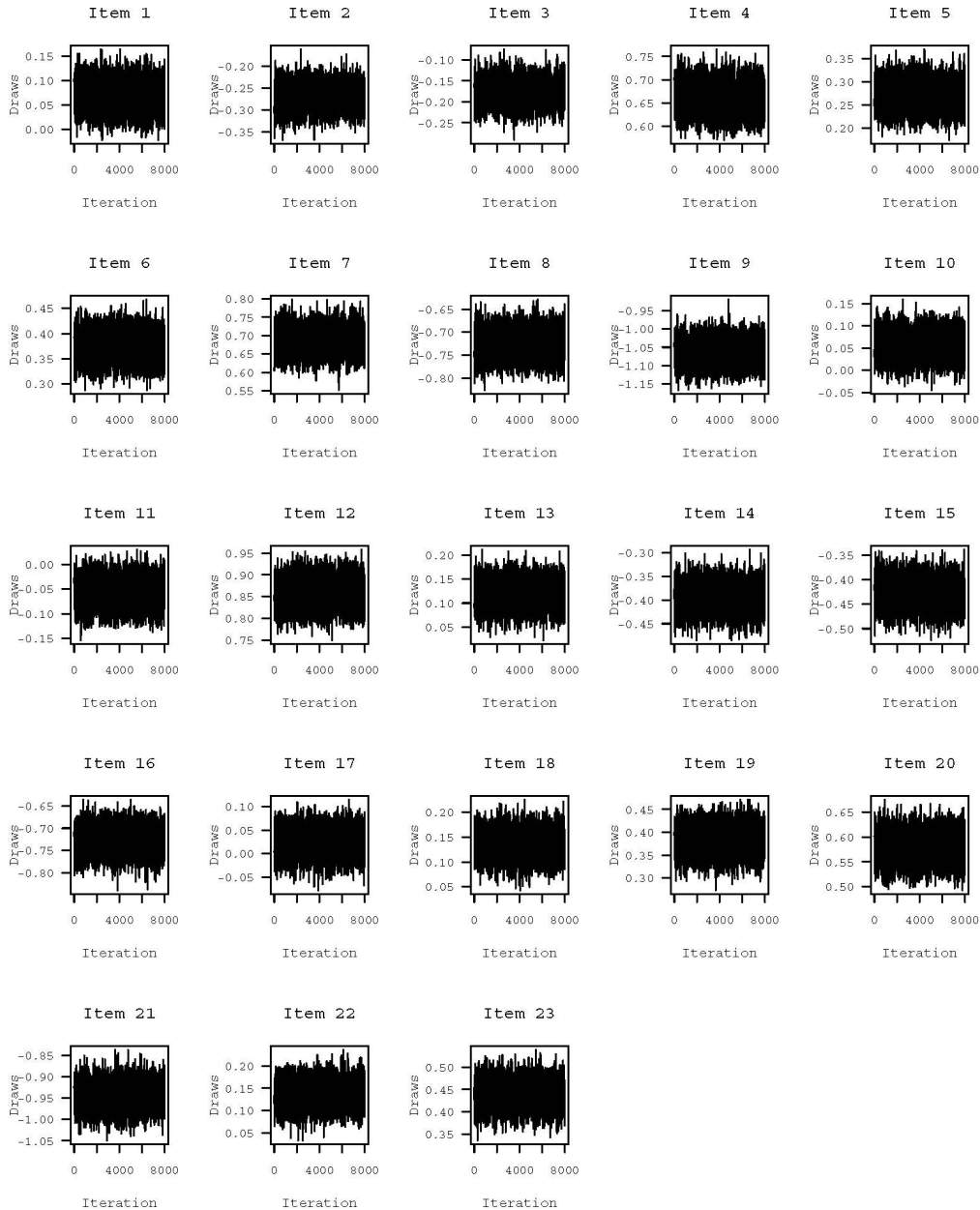

Figure 3:
Trace plots for the difficulty parameters ($\xi_j, j = 1, \ldots, J$)

Table 4:
Mathematical competencies: Parameter estimates of random coefficient IRT model

	Non-parametric imputation N=3615			Complete cases N=2955		
Parameter	mean	sd	95% HDI	mean	sd	95% HDI
γ1 (constant)	−1.202	0.071	[-1.344;-1.062]	−1.183	0.082	[-1.344;-1.027]
γ2 (female)	−0.098	0.017	[-0.131;-0.064]	−0.112	0.019	[-0.150;-0.075]
γ3 (nongerman)	−0.156	0.026	[-0.209;-0.105]	−0.189	0.033	[-0.256;-0.123]
γ4 (selfconcept)	0.147	0.010	[0.127;0.168]	0.154	0.012	[0.131;0.177]
γ5 (schoolsat)	0.016	0.003	[0.009;0.022]	0.017	0.004	[0.009;0.024]
γ6 (peduc)	0.054	0.004	[0.047;0.061]	0.052	0.004	[0.044;0.060]
γ7 (nobooks1)	0.066	0.051	[-0.034;0.165]	0.079	0.060	[-0.040;0.195]
γ8 (nobooks2)	0.213	0.048	[0.122;0.308]	0.226	0.056	[0.113;0.336]
γ9 (nobooks3)	0.293	0.048	[0.198;0.390]	0.308	0.057	[0.196;0.422]
γ10 (nobooks4)	0.399	0.050	[0.302;0.498]	0.415	0.058	[0.300;0.527]
γ11 (nobooks5)	0.389	0.051	[0.289;0.489]	0.421	0.059	[0.304;0.538]
σ^2	0.162	0.006	[0.151;0.175]	0.166	0.007	[0.153;0.180]
α1	1.142	0.059	[1.025;1.258]	1.116	0.064	[0.989;1.243]
α2	1.199	0.061	[1.082;1.320]	1.240	0.068	[1.109;1.376]
α3	1.057	0.056	[0.951;1.169]	1.045	0.063	[0.922;1.170]
α4	0.789	0.050	[0.691;0.888]	0.822	0.057	[0.712;0.933]
α5	1.229	0.058	[1.116;1.341]	1.192	0.065	[1.068;1.323]
α6	0.752	0.047	[0.661;0.844]	0.768	0.051	[0.669;0.870]
α7	1.374	0.063	[1.253;1.499]	1.397	0.069	[1.263;1.536]
α8	1.050	0.064	[0.928;1.178]	1.075	0.072	[0.936;1.217]
α9	1.271	0.079	[1.121;1.427]	1.223	0.091	[1.046;1.404]
α10	1.399	0.064	[1.276;1.527]	1.322	0.069	[1.190;1.458]
α11	1.201	0.055	[1.095;1.310]	1.179	0.062	[1.060;1.302]
α12	0.662	0.050	[0.563;0.760]	0.687	0.055	[0.581;0.799]
α13	0.674	0.050	[0.575;0.770]	0.713	0.054	[0.609;0.817]
α14	0.934	0.057	[0.822;1.046]	0.895	0.062	[0.773;1.018]
α15	1.120	0.059	[1.004;1.237]	1.085	0.065	[0.959;1.214]
α16	0.997	0.059	[0.883;1.114]	0.982	0.068	[0.847;1.116]
α17	1.019	0.054	[0.914;1.125]	1.064	0.061	[0.944;1.184]
α18	0.880	0.048	[0.787;0.976]	0.889	0.054	[0.782;0.993]
α19	1.050	0.054	[0.946;1.156]	1.015	0.059	[0.901;1.130]
α20	0.748	0.048	][0.654;0.843]	0.763	0.052	[0.659;0.863]
α21	0.795	0.062	[0.672;0.914]	0.789	0.071	[0.651;0.928]
α22	1.122	0.055	[1.015;1.231]	1.081	0.060	[0.965;1.198]
α23	1.092	0.053	[0.990;1.198]	1.155	0.061	[1.035;1.276]
ξ1	0.069	0.027	[0.016;0.121]	0.052	0.030	[-0.005;0.113]
ξ2	−0.270	0.025	[-0.319;-0.220]	−0.246	0.030	[-0.306;-0.187]
ξ3	−0.177	0.026	[-0.228;-0.127]	−0.189	0.030	[-0.246;-0.131]

	Non-parametric imputation N=3615			Complete cases N=2955		
Parameter	mean	sd	95% HDI	mean	sd	95% HDI
ξ4	0.662	0.028	[0.607;0.717]	0.687	0.033	[0.622;0.750]
ξ5	0.269	0.027	[0.217;0.321]	0.253	0.031	[0.193;0.314]
ξ6	0.374	0.025	[0.325;0.424]	0.384	0.029	[0.327;0.441]
ξ7	0.687	0.031	[0.627;0.748]	0.706	0.036	[0.638;0.778]
ξ8	−0.724	0.027	[-0.777;-0.672]	−0.725	0.030	[-0.784;-0.665]
ξ9	−1.062	0.030	[-1.123;-1.002]	−1.082	0.034	[-1.147;-1.016]
ξ10	0.055	0.027	[0.003;0.107]	0.034	0.030	[-0.026;0.093]
ξ11	−0.058	0.025	[-0.107;-0.010]	−0.056	0.029	[-0.113;0.000]
ξ12	0.858	0.029	[0.801;0.915]	0.872	0.034	[0.805;0.938]
ξ13	0.118	0.026	[0.068;0.167]	0.145	0.029	[0.087;0.202]
ξ14	−0.394	0.026	[-0.446;-0.343]	−0.425	0.030	[-0.484;-0.367]
ξ15	−0.428	0.026	[-0.478;-0.378]	−0.438	0.029	[-0.494;-0.382]
ξ16	−0.728	0.026	[-0.780;-0.679]	−0.769	0.030	[-0.827;-0.710]
ξ17	0.024	0.025	[-0.025;0.074]	0.036	0.029	[-0.021;0.094]
ξ18	0.134	0.024	[0.087;0.182]	0.153	0.028	[0.098;0.209]
ξ19	0.383	0.027	[0.331;0.435]	0.375	0.031	[0.313;0.436]
ξ20	0.580	0.026	[0.529;0.631]	0.602	0.030	[0.542;0.661]
ξ21	−0.945	0.028	[-0.999;-0.891]	−0.962	0.032	[-1.027;-0.900]
ξ22	0.138	0.026	[0.087;0.188]	0.125	0.029	[0.068;0.182]
ξ23	0.439	0.027	[0.387;0.491]	0.468	0.032	[0.406;0.530]

4 Conclusion

In large scale assessments researchers are usually interested in explaining competence scores by individual characteristics and context variables. Measurement error in competence scores as well as missing values in background variables capturing individual characteristics and context variables need to be accounted for. We propose a Bayesian data augmented MCMC approach that simultaneously estimates plausible values and accounts for missing values in background variables. With this approach latent relationships between competence scores and background variables may be estimated, which efficiently incorporate the uncertainty stemming from only partially observed background variables into parameter estimation. Especially the iterative use of updated parameter values from posterior sampling within the full conditional distribution of missing values is an appealing feature of our approach. Treatment of missing values in background variables using the device of data augmentation further has the advantage that, in contrast to approaches using dummy indicators to model missingness of values, the interpretability of parameters governing the relationship between background variables and competencies remains the same as in situations with completely observed background variables. In a simulation study the proposed approach shows high statistical

accuracy as well as the ability to adequately recover the model parameters, even when higher rates of missing values occur in the data. The applicability to educational large scale research data has been illustrated via an empirical example.

So far, in large scale assessments the issue of missing values in background variables of IRT models has been typically approached by a two-step procedure, first accounting for missing values in background variables using indicator variables and then estimating plausible values. With this procedure, the latent competence is not efficiently included in the imputation of the background variables and the uncertainty stemming from the imputation is not directly accounted for. The proposed approach in this paper allows for simultaneously accounting for measurement error and missing values in background variables and, thus, efficiently incorporates both sources of uncertainty in the parameter and variance estimation. This approach is not only applicable to competence measurement in large scale studies, but to any study in which relationships are estimated for latent constructs with explaining variables that include missing values. The evidence from the simulation study documents that a moderate sample size of 1000 and 10 measurements per individual facilitate accurate parameter estimation, however, given the high accuracy, even smaller sample sizes may suffice.

Future research could focus on extending the suggested estimation approach towards other model frameworks and capturing the different demands of large scale assessment studies. Extending the proposed approach to incorporate other IRT models would strengthen its applicability in empirical studies. As many large scale studies apply a two-stage sampling scheme, drawing first from a set of schools and then from a set of students within these schools, incorporation of hierarchical structures within our estimation approach would also extend the applicability of the suggested approach. Furthermore, often more than one competence is of interest or change in competence scores is considered. In future research the proposed approach could be extended to multidimensional models including both within and between multidimensionality.

Acknowledgements

This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 3 '- 5th grader (Schule, Ausbildung und Beruf), doi:10.5157/NEPS:SC3:1.0.0. From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network. Further, we thank the anonymous referees for the very constructive comments, which helped to improve this paper. The authors also gratefully acknowledge financial support by the Deutsche Forschungsgemeinschaft (DFG) for their research project (“Analyzing relations between latent competencies and context information in the National Educational Panel Study”) within the DFG priority programme 1646 (“Education as a Lifelong process”) under grants AS 368/3-1, PO 1655/2-1, and CA 289/8-1.

References

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

Allen, N., Carlson, J., Johnson, E., & Mislevy, R. (2001). *Scaling procedures.* U.S. Department of Education: The NAEP Technical Report.

Aßmann, C., & Boysen-Hofgrefe, J. (2011). A bayesian approach to model-based clustering for binary panel probit models. *Computational Statistics & Data Analysis*, *55,* 261-279.

Aßmann, C., Gaasch, C., Pohl, S., & Carstensen, C. (2016). Estimation of plausible values considering partially missing backround information: A data augmented MCMC approach. In H.-P. Blossfeld, J. von Maurice, J. Skopek, & M. Bayer (Eds.), *Methological Issues of Longitudinal Surveys* (pp. 505-522).Wiesbaden: Springer.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC Estimation and some Model-Fit Analysis of Multidimensional IRT Models. *Psychometrika, 66,* 541-562.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (S. 397-479). Reading, MA: Addison-Wesley.

Blossfeld, H.-P., Roßnach, H.-G., & Maurice, J. (2011). Education as a Lifelong Process. The German National Educational Panel Study (NEPS) [Special Issue]. *Zeitschrift für Erziehungswissenschaften, 14, 283-299*. Wiesbaden: VS Verlag für Sozialwissenschaften Springer.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression trees.* Chapman & Hall.

Burgette, L., & Reiter, J. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, *172,* 1070-1076.

Chib, S. (2001). Markov Chain Monte Carlo Methods: Computation and inference. *Handbook of Econometrics*, *5,* 3569-3649.

Doove, L., van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, *72,* 92-104.

Duchhardt, C., & Gerdes, A. (2012). *Neps technical report for mathematics: Scaling results of starting cohort 3 in fifth grade.* University of Bamberg: Leibniz Institute for Educational Trajectories.

Edwards, M. (2010). A Markov Chain Monte Carlo Approach to Confirmatory Item Factor Analysis. *Psychometrika*, *75,* 474-497.

Koskin, J., Robins, G., & Pattison, P. (2010). Analysing exponential random graph (p-star) models with missing data using bayesian data augmentation. *Statistical Methodology, 7,* 366-384.

Liu, M., Taylor, J., G., M., & Belin, T. (2000). Multiple Imputation and Posterior Simulation for Multivariate Missing Data in Longitudinal Studies. *Biometrics*, *4,* 1157-1163; 4(56).

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

Mislevy, R. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement*, *11,* 81-91.

Mullis, I., Martin, M., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics.* Bosten College: MA: TIMSS & PIRLS International Study Center.

National Center for Education Statistics. (2013). *The nation's report card: A first look: 2013 mathematics and reading (NCES 2014-451).* Washington D.C.: Institution of Education Sciences, U.S. Department of Education.

Neumann, I., Durchhardt, C., Ehmke, T., Grüßing, M. H., & Knopp, E. (2013). Modeling and assessing of mathematical competence over the lifespan. *Journal for Educational Research Online*, *5,* 80-109.

OECD. (2012). *PISA 2009 Technical Report.* OECD Publishing.

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

Pohl, S., & Carstensen, C. (2012). *Scaling the data of the competence tests. NEPS Technical Report 14.* Bamberg: Otto-Friedrich-University, Nationales Bildungspanel.

Pohl, S., Gräf, 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.

Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests.* Danish Institute for Educational Research.

Rubin, D. (1987). *Multiple imputation for Nonresponse in Surveys.* J. Wiley & Sons.

Schafer, J. (1999). Multiple Imputation: a primer. *Statistical Methods in Medical Research*, *8,* 3-15.

Skopek, J., Pink, S., & Bela, D. (2013). *Starting cohort 3: Grade 5 (sc3). suf version 1.0.0. data manual (neps research data paper).* University of Bamberg, Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Tanner, M., & Wong, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of American Statistical Association*, *92,* 528-540.

van Buuren, S., & Groothuis-Oudshoorn, K. (1987). mice: Multivariate imputation by chain equations. *Journal of Statistical Software*, *45,* 1-67.

Weinert, S., Artel, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C. (2011). Development of competencies across the life span. *Education as a Lifelong Process. The German National Educational Panel Study (NEPS)*, S. 67-86.