Psychological Test and Assessment Modeling, Volume 57, 2015 (4), 472-498

Commonalities and differences in IRT-based methods for nonignorable item nonresponses

Norman Rose[1], *Matthias von Davier*[2] *& Benjamin Nagengast*[3]

Abstract

Missing responses resulting from omitted or not-reached items are beyond researchers' control and potentially threaten the validity of test results. Empirical evidence concerning the relationship between missingness and test takers' performance on the test have suggested that the missing data mechanism is nonignorable and needs to be taken into account. Various IRT-based models for nonignorable item nonresponses have been proposed (Glas & Pimentel, 2008; Holman & Glas, 2005; Korobko, Glas, Bosker, & Luyten, 2008; Moustaki & Knott, 2000; O'Muircheartaigh & Moustaki, 1999; Rose, 2013; Rose, von Davier, & Xu, 2010). In this article, we adopted Rubin's (1976) definitions of missing data mechanisms for educational and psychological measurement and consider the implications for maximum likelihood (ML) estimation in IRT models for incomplete data. Next, we derived multidimensional IRT models for nonignorable item nonresponses. Further, we investigated latent regression models and multiple group IRT models for nonignorable missing responses and compared to multidimensional IRT models. Although these models have a great deal in common, there are important distinctions in the underlying assumptions and restrictions; these have critical implications with respect to their use in real applications. Then, we provided additional insight on how models for nonignorable item nonresponses adjust for missing responses. Finally, we offered a list of guiding questions, which support the choice of appropriate models in concrete applications.

Keywords: Nonignorable, item nonresponses, omitted and not-reached items, IRT

[1] *Correspondence concerning this article should be addressed to:* Norman Rose, PhD, Hector Research Institute of Education Sciences and Psychology, Europastr. 6, 72072 Tübingen, Germany; email: norman.rose@uni-tuebingen.de

[2] Educational Testing Service

[3] Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Germany

Missing data are an inevitable problem for applied researchers. They may occur for many different reasons. For example, participants may not be willing to participate in a study, leading to unit nonresponses, or participants may be unable or unwilling to answer all items of a test. Such item nonresponses typically result from omitted or not-reached items and are common in educational assessments. Furthermore, test takers provide answers that cannot be scored meaningfully, producing item nonresponses due to not-codable item responses. Unplanned missing data resulting from test takers' response behavior must be distinguished from planned missing data due to the design (Graham, Taylor, & Cumsille, 2001; Graham, Taylor, Olchowski, & Cumsille, 2006). Especially in large scale assessments (LSA), only subsets of items are assigned to test takers to reduce costs, participant burden, fatigue, or potential practice effects. With an appropriate test design, including randomized assignment of the different test forms, planned missing data does not pose a threat to validity. Therefore, we focus on the more challenging case of unplanned missing data, which pose not only a loss of efficiency, but potentially lead to biased estimation of item and person parameters in the measurement model. In large scale assessments (LSA), parameters of the structural model such as means, variances, covariances of latent variables are of primary interest instead of individual proficiency levels; however, these distributional parameters may also be biased due to item nonresponses.

Many different approaches to handle missing values have been proposed. Weighting methods, such as inverse probability weighting, are commonly applied to account for unit nonresponses (Li, Shen, Li, & Robins, 2011; Wooldridge, 2007). The simplest approach for item nonresponses is listwise deletion, the inclusion of complete cases into the statistical analysis. Pairwise deletion was proposed as an alternative for models that are based on bivariate statistics, such as structural equation models (SEM) that use covariance matrices as input for parameter estimation. Single and multiple imputation methods rest upon the idea that one should replace missing values with predicted or plausible values in the first step (imputation phase). Next, the augmented data sets are analyzed with standard methods in the second step (analysis phase). In contrast, model-based approaches, such as full information maximum likelihood (FIML), allow for parameter estimation with incomplete data sets. The suitability of the different missing data handling methods depend on whether certain assumptions hold. These assumptions can be derived from Rubin's taxonomy of missing data (1976; 2002). He distinguishes between three missing data mechanisms: Missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). We will examine these mechanisms in greater detail later in this paper. So far it suffices to note that missing data that are MCAR and MAR are also called *ignorable*. In this case, missingness is either completely independent of the observed and unobserved variables under examination (MCAR), or conditionally stochastically independent of the unobserved variables given the observed variables (MAR). The stochastic independencies imply that missingness is not informative with respect to unobserved variables and underlying model parameters and can therefore be ignored. Almost all modern missing data methods rest upon the assumption that the missing data mechanism is ignorable. This is also true for methods like FIML and multiple imputation, which are regarded as state of the art methods for item nonresponses (Schafer & Graham, 2002). In contrast, missing data that are NMAR are termed nonignorable. In this case, missingness is not conditionally independent of the unobserved variables given the observed variables. Such missingness is also called informative with respect to unobserved variables. Discarding this information may result in a biased estimation of model parameters.

In order to account for nonignorable missing data, two classes of models, selection models (SLM; Heckman, 1976, 1979) and pattern mixture models (PMM; Little, 1993, 2008) have been developed. The rationale underlying both models is to combine a model of missingness and the analytic model of substantial interest. Information about missingness in this joint model is used to adjust for nonresponses in variables of the analytic model. Such approaches are of particular interest for item nonresponses in low-stakes educational assessments since evidence has been repeatedly found suggesting that missingness due to omitted and not-reached items depends on the persons' performance in the test. Rose et al. (2010) reported a negative correlation ($r = -.330$) between the response rate and the proportion correct score in the PISA 2006 data. They also found that easier items were more often completed than difficult items. Test takers do not seem to omit items randomly. Pohl, Gräfe, and Rose (2014) also found negative correlations between the latent ability and (a) the propensity to omit items ($r = -.175$), and (b) the propensity for item nonresponses due to not-reached items ($r = -.200$). Similarly, Culbertson (2011) found higher rates of omitted items in persons with lower test scores, at least in items where students are required to produce a response actively. Since test takers' performance on a test is an indicator of their latent abilities, which are inherently unobservable, the MAR assumption seems questionable.

Starting with O'Muircheartaigh and Moustaki (1999) and Moustaki and Knott (2000), multidimensional IRT models for nonignorable missing responses have been proposed and further developed in a series of papers (Glas & Pimentel, 2008; Holman & Glas, 2005; Korobko et al., 2008; Moustaki & Knott, 2000; O'Muircheartaigh & Moustaki, 1999; Rose, 2013; Rose, Von Davier, & Nagengast, 2015; Rose et al., 2010). It could be demonstrated that these models are special cases of both PMM (O'Muircheartaigh & Moustaki, 1999) and SLM (Rose, 2013) adapted for the case of latent trait models under specific assumptions. Different MIRT models for missing responses have been discussed in the literature. In all these MIRT models a latent response propensity is assumed that underlies the response indicator variables which represent missingness of item responses. This approach is promising for many real applications but suffers from increased model complexity and strong assumptions. Rose et al. (2010) and Rose (2013) proposed latent regression models (LRM) and multiple group (MG) IRT models for nonignorable item nonresponses as alternatives, which are computationally less demanding and more flexible. Although closely related to MIRT models, the LRM and the MG-IRT model rest upon different assumptions. Therefore, these different models may be more or less appropriate in different applications.

The aim of this paper is threefold: First, we clarify the underlying rationale of IRT models for nonignorable item nonresponses by deriving these models from traditional SLMs. Second, we compare MIRT models, LRMs, and MG-IRT models for missing responses and reveal the commonalities and differences. We explain the assumptions of the different approaches, highlight the advantages and disadvantages of each model, and discuss the implications for their use in real applications. Third, we provide guidance regarding the choice between the different model-based approaches and their application to real data.

This paper is organized as follows: We begin with definitions of missing data mechanisms following Rubin's taxonomy, but with some adoptions for peculiarities of educational and psychological measurement. We will then consider ML estimation in IRT models with the focus on marginal ML (MML) estimation as the most commonly used estimator for MIRT models. We will then examine the relationship of MIRT models to the LRMs and MG-IRT models. Subsequently, we provide some insights in how these models adjust for missing responses considering the EM algorithm used for MML estimation. After a synoptic comparison of the different models, the implications of the theoretical considerations with respect to applications will be discussed. In order to support selection and application of models for nonignorable item nonresponses, we also provide a list of relevant questions that can serve as a guide for applied researchers. We conclude with a summary and suggestions for future research.

Defining the missing data mechanism in educational and psychological measurement

The phrase missing data mechanism does not imply causality, rather the three missing data mechanisms considered here are defined by stochastic dependencies between observed and unobserved variables (Schafer & Graham, 2002). Following Rubin (1976), the observed data matrix $\boldsymbol{Y} = \boldsymbol{y}$ can be decomposed into an observed part $\boldsymbol{Y_o} = \boldsymbol{y_o}$ and a missing part $\boldsymbol{Y_m} = \boldsymbol{y_m}$ so that $\boldsymbol{Y} = (\boldsymbol{Y_o}, \boldsymbol{Y_m})$ and $\boldsymbol{y} = (\boldsymbol{y_o}, \boldsymbol{y_m})$. Missingness is conceptualized as a random variable within the same probability space like $\boldsymbol{Y}$. In his original formulation Rubin did not distinguish between dependent or independent variables, auxiliary variables, and covariates. However, in psychological and educational measurement, the items constituting a measurement model, the covariates, such as gender, socioeconomic status, motivation, etc., and the latent variables are different types of variables. During the scaling procedures in application of measurement models these variables are treated differently. Accordingly, we consider these distinct groups of variables in the definitions of the missing data mechanisms.

Let N be the sample size with the observations $n = 1, \ldots, N$. Furthermore, the measurement model of potentially multidimensional latent variable $\boldsymbol{\xi}$ is constituted by items $i = 1, \ldots, I$. Let $\boldsymbol{Y}$ be a $N \times I$ random matrix consisting of binary random variables Y_{ni} that code the item responses. In achievement tests that may simply be an incorrect ($Y_{ni} = 0$) or a correct response ($Y_{ni} = 1$). The manifest variables Y_{ni} are fallible measures of a k-dimensional latent variable $\boldsymbol{\xi} = \xi_1, \ldots, \xi_k$. The corresponding response indicator variables D_{ni} indicate the observational status of Y_{ni}, where

$$D_{ni} = \begin{cases} 1, \text{ if } Y_{ni} \text{ is observed} \\ 0, \text{ if } Y_{ni} \text{ is not observed} \end{cases} \tag{1}$$

The matrix $\boldsymbol{D}$ is a $N \times I$ matrix with the elements D_{ni}. All manifest variables Z_{nj}, with j = 1,..., J, that are not part of the measurement model of $\boldsymbol{\xi}$ constitute the covariate vector $\boldsymbol{Z_n}$ for case n. Accordingly, $\boldsymbol{Z}$ is a $N \times J$ matrix. Here we assume that the covariates are fully observed. Hence, the observed part of the variables in the target model is now ($\boldsymbol{Y_o}$,$\boldsymbol{Z}$) and the missing part ($\boldsymbol{Y_m}$,ξ). Furthermore, the response indicator matrix $\boldsymbol{D}$ is fully observed, but typically is not part of the analytical model. Following Rubin (1976) we can define the missing data mechanisms MCAR, MAR, and NMAR. Although essentially equivalent, there exist different definitions of these missing data mechanisms (e. g. Schafer & Graham, 2002; Seaman, Galati, Jackson, Carlin, et al., 2013). It is crucial to remember that our definitions are based on probability statements about random variables with a joint distribution. Observed data are realizations of random variables. Therefore, the missing data mechanisms describe stochastic data generating processes in terms of probability statements. As the definitions are based on probability theory, we can deal with events that may never be observable, such as the probability of solving an omitted or not-reached item. Accordingly, $\boldsymbol{Y_m}$ does not refer to realized but unobserved item responses. The underlying rationale is that test takers could have solved omitted or not reached items even if they had completed these items. We could then observe the realizations $\boldsymbol{y_m}$ (in other words, we took a counterfactual perspective on item nonresponses).

The missing data mechanism with respect to $\boldsymbol{Y}$ is denoted as MCAR if

$$\boldsymbol{D} \perp (\boldsymbol{Y}, \boldsymbol{Z}, \xi)\,. \tag{2}$$

Thus, missingness is stochastically independent of the latent and manifest variables of the measurement model as well as the covariates in the model. Due to the distinction between items responses $\boldsymbol{Y}$ and covariates $\boldsymbol{Z}$, three different MAR conditions result. Each MAR condition implies a different method in order to adjust for item nonresponses. The missing data mechanism with respect to $\boldsymbol{Y}$ is called to be MAR given $\boldsymbol{Z}$ if

$$\boldsymbol{D} \perp \left(\boldsymbol{Y}, \xi\right) \mid \boldsymbol{Z} \tag{3}$$

In this case missingness is conditionally stochastically independent of the unobserved and observed variables in the model given the covariates. However, unconditional stochastic dependency between $\boldsymbol{D}$ and ($\boldsymbol{Y_m}$, ξ) does not contradict with the definition given by Equation 3. For example, in two stage testing designs with an incomplete matrix design a routing test ($\boldsymbol{Z}$) may be used to decide which test form and, therefore, which items will be assigned in the final test ($\boldsymbol{Y}$). An appropriate routing test should be substantially related to the latent ability ξ. However, the missing pattern depends on the test form which is determined by $\boldsymbol{Z}$, implying that Equation 3 holds. The practical implication is that the covariates need to be included as auxiliary variables for parameter estimation, even if they are not of substantial interest.

The second MAR assumption refers to the situation where missingness depends only on observed item responses. Accordingly, the missing data mechanisms with respect to $\boldsymbol{Y}$ is called MAR given $\boldsymbol{Y_o}$ if

$$\boldsymbol{D} \perp \left(\boldsymbol{Y}_m, \xi\right) \mid \boldsymbol{Y}_o \tag{4}$$

The $\boldsymbol{Y_o}$-conditional independence of missingness off the unobserved variables hold in computerized adaptive testing if either the starting item is fixed or randomly selected from the item pool (Glas, 2006; Mislevy & Wu, 1996). In this case missing responses result from not administering items in individual assessments. The missing pattern is determined by the item selection algorithm, which does not include covariates or any unobserved variables. It is worth noting that the assumption of MAR given $\boldsymbol{Y_o}$ is implicitly made in most applications of IRT models with missing data.

The most general case in which the missing data mechanism is ignorable is the independence of missingness off any unobserved variables in the model given both types of observable variables: (a) the covariates, and (b) the observed item responses. Hence, the missing data mechanism with respect to $\boldsymbol{Y}$ is called MAR given ($\boldsymbol{Z}$, $\boldsymbol{Y_o}$) if Equations 3 and 4 do not hold and conditional stochastic independence

$$\boldsymbol{D} \perp (\boldsymbol{Y_m}, \xi) \mid (\boldsymbol{Y_o}, \boldsymbol{Z}) \tag{5}$$

is valid. Note that $\boldsymbol{Z}$-conditional independence (see Equation 3) and $\boldsymbol{Y_o}$-conditional independence (see Equation 4) are special cases of ($\boldsymbol{Z}$, $\boldsymbol{Y_o}$)-conditional independence of missingness off the unobserved variables. However, they are not equal. In an application of computerized adaptive testing, for example, where the choice of starting items depends on background variables ($\boldsymbol{Z}$), such as educational or professional qualifications, the selection of items and, the missing data pattern due to not-administered items is determined not only by the observed item responses, but from covariates as well (Glas, 2006). Furthermore, the assumption of MAR given ($\boldsymbol{Z}$, $\boldsymbol{Y_o}$) is commonly made in most IRT-based scaling procedures with incomplete item responses and additional covariates included in a background model. In almost all educational LSA, such as Programme for International Student Assessment (PISA; OECD, 2009) or Trends in International Mathematics and Science Study (TIMSS; Adams et al., 1998), covariates are included in the item and person parameter estimation.

Finally, the missing data mechanism is called not missing at random if

$$\boldsymbol{D} \not\perp (\boldsymbol{Y_m}, \xi) \mid (\boldsymbol{Y_o}, \boldsymbol{Z}) \tag{6}$$

Here $\not\perp$ indicates stochastic dependency. Thus, the missing data are nonignorable if the missingness and unobserved variables in the model are conditionally stochastically dependent on each other given the observed variables in the model. For example, if there is any subpopulation defined by the value $\boldsymbol{Z} = \boldsymbol{z}$ and the observed response pattern $\boldsymbol{Y_o} = \boldsymbol{y_o}$ in which missingness in $\boldsymbol{Y}$ depends on the latent variable or missing items, then the item nonresponses are NMAR.

ML estimation in IRT models with item nonresponses

Rubin's taxonomy of missing data is very useful not only for classification, but can be used to derive appropriate methods to adjust for item nonresponses. The unconditional and conditional independence assumptions defining the missing data mechanisms can be applied to the estimators of interest. Conclusions can be drawn about how to account for missingness appropriately. In this article, we restrain our focus to ML estimation in IRT measurement models. In the case of missing responses it is important to differentiate between the complete data likelihood $L(\boldsymbol{Y}, \boldsymbol{D} \mid \boldsymbol{Z}; \boldsymbol{\iota}, \boldsymbol{\varphi})$ and the observed data likelihood $L(\boldsymbol{Y}_o, \boldsymbol{D} \mid \boldsymbol{Z}; \boldsymbol{\iota}, \boldsymbol{\varphi})$ (Little & Rubin, 2002; Schafer, 1997). The parameter vector $\boldsymbol{\iota}$ refers to the measurement model of ξ and includes the item parameters, parameters of a potential latent regression model and so on. The vector $\boldsymbol{\varphi}$ consists of parameters of the missingness model, for example, parameters of probit regressions $P(D_{ni} = 1 \mid \boldsymbol{Z}_n, \boldsymbol{Y}_n)$ that are included in SLM for normally distributed variables (Heckman, 1976, 1979). As common in many IRT models, the multidimensional covariate $\boldsymbol{Z}$ is taken into account as a purely exogenous variable and is required to be fully observed respectively. In ML estimation the complete data likelihood is proportional to the joint distribution of the variables $P(\boldsymbol{Y}, \boldsymbol{D} \mid \boldsymbol{Z})$ given the exogenous variables in the model. Note that $P(.)$ denotes either probabilities or probability mass functions for discrete random variables. Probability density functions of continuous random variables are denoted by $g(.)$. Using this notation, the complete data likelihood for dichotomous items can be written as

$$L(\boldsymbol{Y}, \boldsymbol{D} \mid \boldsymbol{Z}; \boldsymbol{\iota}, \boldsymbol{\varphi}) \propto P(\boldsymbol{Y}, \boldsymbol{D} \mid \boldsymbol{Z}; \boldsymbol{\iota}, \boldsymbol{\varphi}) \tag{7}$$

There are different ML estimators for IRT models, such as joint maximum likelihood (JML) estimation, marginal maximum likelihood estimation (MML) for one and two-parameter models, and conditional maximum likelihood (CML) estimation for unidimensional and multidimensional Rasch models only (e. g., Baker & Kim, 2004). In the following, we focus on MML estimation since consistency of JML estimates is not ensured (Drasgow, 1989; Lord, 1986) and CML estimation is limited to one-parameter models. Therefore, MML estimation is currently the most common method to obtain estimates of one, two, and three-parameter IRT models. MML estimation includes a model of the latent distribution $g(\xi \mid \boldsymbol{Z})$. Typically a parametric distribution is assumed, such as the conditional normal distribution of ξ given $\boldsymbol{Z}$. Instead of the individual values of ξ, the parameters of this distribution, such as means and variances, are estimands in the parameter vector $\boldsymbol{\iota}$. The complete data likelihood for N independent observations under the MML paradigm is

$$L(\boldsymbol{Y}, \boldsymbol{D} \mid \boldsymbol{Z}; \boldsymbol{\iota}, \boldsymbol{\varphi}) \propto \int_{N \times \mathbb{R}^k} g(\boldsymbol{Y}, \boldsymbol{D}, \xi \mid \boldsymbol{Z}; \boldsymbol{\iota}, \boldsymbol{\varphi})\, d\xi \tag{8}$$

$$L(\boldsymbol{Y}, \boldsymbol{D} \mid \boldsymbol{Z}; \boldsymbol{\iota}, \boldsymbol{\varphi}) \propto \prod_{n=1}^{N} \int_{\mathbb{R}^k} g(\boldsymbol{Y}_n, \boldsymbol{D}_n, \xi \mid \boldsymbol{Z}_n; \boldsymbol{\iota}, \boldsymbol{\varphi})\, d\xi. \tag{9}$$

Since only the distribution of the latent variable ξ is of interest in MML estimation and not individual values ξ_n, the subscript n is omitted from all latent variables in MML estimation equations. The joint distribution of the right-hand side of the equation can be factored in different ways. In order to derive a general MML estimator of the measurement model of ξ considering missingness we use

$$L(\boldsymbol{Y},\boldsymbol{D}\mid\boldsymbol{Z};\boldsymbol{\iota},\boldsymbol{\varphi})\propto\prod_{n=1}^{N}\int_{\mathbb{R}^k}P(\boldsymbol{D}_n\mid\boldsymbol{Y}_n,\xi,\boldsymbol{Z}_n;\boldsymbol{\varphi})P(\boldsymbol{Y}_n\mid\xi,\boldsymbol{Z}_n;\boldsymbol{\iota})\,g(\xi\mid\boldsymbol{Z}_n;\boldsymbol{\iota})\,d\xi.\tag{10}$$

which refers to a generalization of SLM to the case of latent trait models. As common in IRT models, we assume that no differential item functioning (DIF) exists depending on any covariate ***Z***, and local stochastic independence between all pairs of items *i* and *i'* hold:

$$\boldsymbol{Y}\perp\boldsymbol{Z}\mid\xi\tag{11}$$

$$\forall(i,i'),i\neq i':Y_i\perp Y_{i'}\mid\xi\tag{12}$$

Based on these assumptions the complete data likelihood can be written as

$$L(\boldsymbol{Y},\boldsymbol{D}\mid\boldsymbol{Z};\boldsymbol{\iota},\boldsymbol{\varphi})\propto\prod_{n=1}^{N}\int_{\mathbb{R}^k}P(\boldsymbol{D}_n\mid\boldsymbol{Y}_n,\xi,\boldsymbol{Z}_n;\boldsymbol{\varphi})P(\boldsymbol{Y}_n\mid\xi;\boldsymbol{\iota})\,g(\xi\mid\boldsymbol{Z}_n;\boldsymbol{\iota})\,d\xi.\tag{13}$$

This likelihood function is of theoretical interest but cannot be used for parameter estimation in the presence of item nonresponses. Instead, the observed data likelihood needs to be used, which is the integral of the complete data likelihood over the missing part. Note that the MML estimator given by Equation 10 already includes an integral over the unobserved variable ξ but not the integral over the missing part $\boldsymbol{Y_m}$. Integrating over all unobserved variables finally yields

$$\begin{aligned}L(\boldsymbol{Y_o},\boldsymbol{D}\mid\boldsymbol{Z};\boldsymbol{\iota},\boldsymbol{\varphi})\propto\prod_{n=1}^{N}\int_{\mathbb{R}^k}P\left(\boldsymbol{Y_o}^{(n)}\mid\xi;\boldsymbol{\iota}\right)\int_{\Omega_m}P\left(\boldsymbol{D}_n\mid\boldsymbol{Y_o}^{(n)},\boldsymbol{Y_m}^{(n)},\xi,\boldsymbol{Z}_n;\boldsymbol{\varphi}\right)\\ P\left(\boldsymbol{Y_m}^{(n)}\mid\xi;\boldsymbol{\iota}\right)g(\xi\mid\boldsymbol{Z}_n;\boldsymbol{\iota})\,d\Omega_m d\xi.\end{aligned}\tag{14}$$

The domain Ω_m indicates the set of all possible response patterns $\boldsymbol{Y_m}=\boldsymbol{y_m}$ that could have been observed jointly with $\boldsymbol{Y_o}=\boldsymbol{y_o}$. In the case of binary items, Ω_m contains $I-\sum_{i=1}^{I}D_{ni}$ possible patterns for case *n*. In the appendix, the derivation of the observed from the complete data likelihood is shown in detail. Here we focus on the conditional distribution $P\left(\boldsymbol{D}_n\mid\boldsymbol{Y_o}^{(n)},\boldsymbol{Y_m}^{(n)},\xi,\boldsymbol{Z}_n;\boldsymbol{\varphi}\right)$ of missingness given all other observed and unobserved variables in the model. From the different definitions of the missing data mechanism this conditional distribution and, therefore, the observed data likelihood can further be simplified. For example, if the missing data mechanism with respect to ***Y*** is MCAR, ***D*** does not depend on any variables in the model (see Equation 2). As demonstrated in the appendix, the observed data likelihood is then

$$L(\boldsymbol{Y_o},\boldsymbol{D}\mid\boldsymbol{Z};\boldsymbol{\iota},\boldsymbol{\varphi})\propto P(\boldsymbol{D};\boldsymbol{\varphi})\prod_{n=1}^{N}\int_{\mathbb{R}^k}P\left(\boldsymbol{Y_o}^{(n)}\mid\xi;\boldsymbol{\iota}\right)g(\xi\mid\boldsymbol{Z}_n;\boldsymbol{\iota})\,d\xi.\tag{15}$$

There exist two independent factors in the likelihood. The first factor refers to the model of missingness and the second factor is the common MML estimator of the measurement model of $\boldsymbol{\xi}$ based on the observed item responses. This implies that ML inference is valid even if $\boldsymbol{D}$ is ignored in estimating $\boldsymbol{\iota}$.

If the missing data mechanism with respect to $\boldsymbol{Y}$ is MAR given $\boldsymbol{Y_o}$, $\boldsymbol{Z}$, or both ($\boldsymbol{Y_o}$, $\boldsymbol{Z}$), the observed data likelihood can also be factored into two independent pieces. Considering the most general MAR assumption of a missing data mechanism that is MAR given ($\boldsymbol{Y_o}$, $\boldsymbol{Z}$), then Equation 14 is equal to

$$L\left(\boldsymbol{Y}_{\boldsymbol{o}},\boldsymbol{D}\mid\boldsymbol{Z};\boldsymbol{\iota},\boldsymbol{\varphi}\right)\propto P\left(\boldsymbol{D}\mid\boldsymbol{Y}_{\boldsymbol{o}},\boldsymbol{Z};\boldsymbol{\varphi}\right)\prod_{n=1}^{N}\int_{\mathbb{R}^k}P\left(\boldsymbol{Y}_{\boldsymbol{o}}^{(n)}\mid\xi;\boldsymbol{\iota}\right)g\left(\xi\mid\boldsymbol{Z}_n;\boldsymbol{\iota}\right)d\xi. \quad (16)$$

It is important to note that this factorization applies only if the covariate $\boldsymbol{Z}$ is included in the model. Hence, all covariates that are related to missingness need to be modeled jointly with the measurement model of $\boldsymbol{\xi}$. If there is no DIF depending on $\boldsymbol{Z}$ (see Equation 11) it is sufficient to include the covariates in the structural model or in a latent regression model, which is expressed by the individual conditional distributions $g(\boldsymbol{\xi} \mid \boldsymbol{Z_n}; \boldsymbol{\iota})$ in the Equations 10 - 16. If the missing data mechanism with respect to $\boldsymbol{Y}$ is MAR given $\boldsymbol{Y_o}$ then

$$L\left(\boldsymbol{Y}_{\boldsymbol{o}},\boldsymbol{D}\mid\boldsymbol{Z};\boldsymbol{\iota},\boldsymbol{\varphi}\right)\propto P\left(\boldsymbol{D}\mid\boldsymbol{Y}_{\boldsymbol{o}};\boldsymbol{\varphi}\right)\prod_{n=1}^{N}\int_{\mathbb{R}^k}P\left(\boldsymbol{Y}_{\boldsymbol{o}}^{(n)}\mid\xi;\boldsymbol{\iota}\right)g\left(\xi\mid\boldsymbol{Z}_n;\boldsymbol{\iota}\right)d\xi. \quad (17)$$

Missingness is independent of any covariates in the model. Accordingly, the latter needs not necessarily be included in the model to adjust for item nonresponses.

If the missing data mechanism with respect to $\boldsymbol{Y}$ is NMAR, the Equation 14 cannot be factorized into two independent pieces that refer to independent models with distinct sets of parameters that could be maximized independently. Accordingly, a model of missingness and a measurement model of $\boldsymbol{\xi}$ needs to be estimated simultaneously to adjust for missing data. Unfortunately, a joint model for ($\boldsymbol{Y_o}$, $\boldsymbol{D}$) that refers to the likelihood function given by Equation 14 is not identified without further restrictions.

Multidimensional IRT models for nonignorable missing responses

If the missing data mechanism is nonignorable, missingness is informative with respect to unobserved variables and model parameters underlying these variables. Discarding this information potentially results in biased parameter estimation and invalid inference. The crucial question in models for nonignorable data is how to identify and to specify the model of missingness that refers to the term $P\left(\boldsymbol{D}_n\mid\boldsymbol{Y}_{\boldsymbol{o}}^{(n)},\boldsymbol{Y}_{\boldsymbol{m}}^{(n)},\boldsymbol{\xi},\boldsymbol{Z}_n;\boldsymbol{\varphi}\right)$ in Equation 14.

Logit or probit regression with the response indicator variables D_i as dependent variables are not applicable due to the missing values in $\boldsymbol{Y}$ and the latent variable $\boldsymbol{\xi}$, which is always missing. Model estimation is only possible under certain assumptions. A well-known example is the univariate normal selection model that rests upon the strong assumption of normality (Heckman, 1976, 1979; Puhani, 2000). Multidimensional IRT models for missing responses are identified by assumptions about model-implied dependencies between manifest variables $\boldsymbol{Y}$ and $\boldsymbol{D}$ due to the structural model of latent variables. It is assumed that a latent response propensity $\boldsymbol{\theta}$ exists that underlies the response indicator variables D_i. Therefore, completing a test requires not only the competency (ξ) for processing test items, but also the motivation, willingness, and cognitive speed to do so. These and many other characteristics of the test takers are potential correlates of $\boldsymbol{\theta}$. However, for the sake of clarity, we exclusively use the terms competency, ability, or proficiency for $\boldsymbol{\xi}$, which is intended to be measured by the items of the test.

Conditional stochastic independence is assumed for all Y_i and D_i given the latent variables. That is

$$\forall\left(i=1,\ldots,I\right): Y_i \perp \left(\boldsymbol{Y}_{\text{-}i}, \boldsymbol{D}\right) \mid \xi \tag{18}$$

$$\forall\left(i=1,\ldots,I\right): D_i \perp \left(\boldsymbol{D}_{\text{-}i}, \boldsymbol{Y}\right) \mid \boldsymbol{\theta} \tag{19}$$

Based on this assumption, also denoted as local stochastic independence, Equation 14 can be simplified to the final MML function of the MIRT model for nonignorable item nonresponses

$$L\left(\boldsymbol{Y},\boldsymbol{D} \mid \boldsymbol{Z};\boldsymbol{\iota},\boldsymbol{\varphi}\right) \propto \prod_{n=1}^{N} \int_{\mathbb{R}^k} \int_{\mathbb{R}^p} \prod_{i=1}^{I} P\left(Y_{ni} = y_{ni} \mid \boldsymbol{\xi};\boldsymbol{\iota}\right)^{d_{ni}} P\left(D_{ni} = d_{ni} \mid \boldsymbol{\theta};\boldsymbol{\varphi}\right) g\left(\xi,\boldsymbol{\theta} \mid \boldsymbol{Z}_n;\boldsymbol{\iota}\right) d\xi d\boldsymbol{\theta}. \tag{20}$$

Note that the subscript n is omitted from all latent variables including $\boldsymbol{\theta}$ since individual values $\boldsymbol{\theta}_n$ are not relevant in MML estimation. Note that the model represented by Equation 20 is conceptually close to the MIRT model proposed by Holman and Glas (2005) and O'Muircheartaigh and Moustaki (1999). The only difference is that covariates $\boldsymbol{Z}$ have been included here. In the final model $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is a p-dimensional latent variable, with $p \geq 1$. Furthermore, the conditional distribution $g(\boldsymbol{\xi}, \boldsymbol{\theta} \mid \boldsymbol{Z}_{\mathrm{n}}; \boldsymbol{\iota})$ refers to the joint distribution of ξ and $\boldsymbol{\theta}$ given the covariates $\boldsymbol{Z}$. The latter does not need to be part of the model in order to adjust for nonignorable missing data, but can be included in the background model to facilitate item parameter estimation or to improve person parameter estimation (Mislevy, 1987, 1988). When applied, the term $\mathrm{g}(\boldsymbol{\xi}, \boldsymbol{\theta} \mid \boldsymbol{Z}; \iota)$ refers to the potentially multivariate latent regression models $E(\xi \mid \boldsymbol{Z})$ and $E(\boldsymbol{\theta} \mid \boldsymbol{Z})$. It is important to note that the validity of the model assumptions (see Equations 18 and 19) implies that all stochastic dependency between missingness ($\boldsymbol{D}$) and the unobserved variables ξ and $\boldsymbol{Y_m}$ result from the relationship between the latent variable ξ and $\boldsymbol{\theta}$. Hence, if ξ and $\boldsymbol{\theta}$ are stochastically independent then $\boldsymbol{Y}$ and $\boldsymbol{D}$ are necessarily independent and the missing data mechanism is MCAR. Similarly, if $\xi \not\perp \boldsymbol{\theta}$ but $\xi \perp \boldsymbol{\theta} \mid \boldsymbol{Z}$, then the missing data mechanism is MAR given $\boldsymbol{Z}$. In other words MIRT models can also be used to study the missing data mechanism of item nonresponses (Pohl et al., 2014).

Figure 1 shows an hypothetical example of a MIRT model, which is mathematically represented by Equation 20. Note that the model allows for within-item multidimensionality within the measurement models of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$. That is, a single item Y_i can be an indicator of more than one latent dimension ξ_m. Similarly, single response indicators D_i can indicate more than one latent propensity θ_l. However, no direct effects of $\boldsymbol{\xi}$ on $\boldsymbol{D}$ or $\boldsymbol{\theta}$ on $\boldsymbol{Y}$ are allowed. Alternative MIRT models have been proposed allowing for additional direct effects across the measurement models of the latent variables (Holman & Glas, 2005; Rose, 2013; Rose et al., 2010). In covariance structure analyses, equivalence of alternative models is typically defined by (a) the equality of model implied covariance and mean structures, and (b) the equality of model fit (Raykov & Penev, 1999; Raykov & Marcoulides, 2001; Stelzl, 1986). In the case of missing data models, further aspects should be considered. These specific models aim to adjust for item nonresponses without changing the target model of substantial interest. The latter is the measurement model of $\boldsymbol{\xi}$. From a practical point of view, two alternative models for missing responses are equivalent in the sense of being interchangeable, if they equally adjust for missing responses. Taking these aspects into account, we consider two missing data models A and B as equivalent if (a) the latent ability variable $\boldsymbol{\xi}$ is equally constructed in Models A and B, (b) the bias of item and person parameters due to missing responses is equally reduced in both models, and (c) both models have the same model fit in terms of any goodness-of-fit statistics. The equality of the construction of $\boldsymbol{\xi}$ is a defining feature of models for item nonresponses, and means that the measurement model of $\boldsymbol{\xi}$ based on $\boldsymbol{Y}$ is preserved as a submodel in the joint missing data model of ($\boldsymbol{Y}$, $\boldsymbol{D}$). Only then can a missing data model correct for nonresponse biases in the parameters of the target model ($\boldsymbol{\iota}$) instead of estimating completely different parameters of a model with a different meaning.

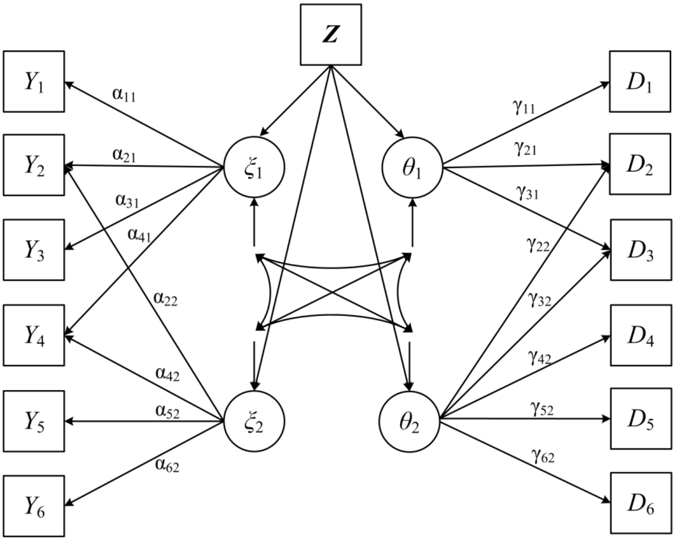

Figure 1:
The MIRT model with multidimensional latent variables $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$

Rose (2013) examined equivalence of different MIRT models analytically. He showed that these models differ in the construction of the latent variables apart from $\boldsymbol{\xi}$. More precisely, the latent response propensity $\boldsymbol{\theta}$ is replaced by functions $f(\boldsymbol{\xi}, \boldsymbol{\theta})$ such as latent difference variables

$$f\left(\xi,\boldsymbol{\theta}\right)=\boldsymbol{\theta}-\sum_{m=1}^{k}\xi_{m} \tag{21}$$

or latent residuals

$$f\left(\xi,\boldsymbol{\theta}\right)=\boldsymbol{\theta}-E\left(\boldsymbol{\theta}\mid\xi\right) \tag{22}$$

Here we abstain from a detailed description of all these alternative MIRT models for two reasons. First, all these alternative models are equivalent to the examined model (see Equation 20) in terms of the construction of ξ, the bias adjustment, and the model fit. Second, the model specification of the alternative MIRT becomes intricate in cases of high dimensional latent variables $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$. Readers interested in equivalent MIRT models are referred to Holman and Glas (2005) and (Rose, 2013).

Latent regression models for nonignorable item nonresponses

MIRT models have several advantages. They do not only allow adjusting for noningorable missing responses, but can also be used to study the nonresponse mechanisms by analyzing $\boldsymbol{D}$ as well as the correlational structure between $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$. However, MIRT models also have limitations. The complexity of the model increases substantially. The number of manifest variables doubles due to including the response indicators D_i. The number of latent dimensions increases as well depending on the dimensionality of $\boldsymbol{\theta}$. Using MML estimation this is critical since MIRT models with more than five latent dimensions are still computationally challenging (Asparouhov & Muthén, 2012; Cai, 2010; Schilling & Bock, 2005). If the rates of missing data are low so that little variation is seen in the response indicators, estimation of parameters of the measurement model of $\boldsymbol{\theta}$ may be inaccurate or even fail, if the sample size is small. Finally, under MML estimation with the assumption of a multivariate normal distribution of $(\boldsymbol{\xi}, \boldsymbol{\theta})$ only linear relationships between the latent ability dimensions ξ_m and the latent propensities θ_l are taken into account. If additional covariates are included in a background model of a MIRT model for missing data, then conditional linear dependencies between ξ_m and θ_l given $\boldsymbol{Z}$ are taken into account only.

To overcome these problems Rose et al. (2010) and Rose (2013) proposed LRMs and MG-IRT models for nonignorable item nonresponses. The basic idea is to use functions $f(\boldsymbol{D})$ as independent variables in latent regressions $E[\xi \mid f(\boldsymbol{D})]$ simultaneously estimated with the parameters $\boldsymbol{\iota}$ of the measurement model of $\boldsymbol{\xi}$. As in MIRT models information of missingness is taken into account while model complexity is reduced by excluding the latent variable model of $\boldsymbol{D}$. Nevertheless, the LRM is closely related to MIRT models for nonignorable item nonresponses if the assumptions of the latter hold. If we could observe the latent responses propensity $\boldsymbol{\theta}$ and could include it as a covariate in the model, then the missing data mechanism with respect to $\boldsymbol{Y}$ were MAR given the covariate $\boldsymbol{\theta}$. $\boldsymbol{D}$ could be omitted from the model. However, $\boldsymbol{\theta}$ can never be observed, but proxies of it may be available. For example, in cases of a unidimensional latent response propensity θ underlying $\boldsymbol{D}$, the sum score $S_D = \sum_{i=1}^{I} D_i$ is increasingly correlated with θ if the number of response indicators increases. Model complexity can be substantially reduced using S_D as a function $f(\boldsymbol{D})$ in a LRM instead of modeling $\boldsymbol{D}$. Strictly speaking the assumption of LRMs for missing responses is

$$\boldsymbol{D} \perp (\boldsymbol{Y_m}, \boldsymbol{\xi}) \mid (f(\boldsymbol{D}), \boldsymbol{Y_o}) \tag{23}$$

In informal terms, this equation means that all information of $\boldsymbol{D}$ is preserved in the function $f(\boldsymbol{D})$. Furthermore, local stochastic independence and absence of DIF is assumed with respect to the function $f(\boldsymbol{D})$, so that

$$\forall (i = 1, \ldots, I) : Y_i \perp [\boldsymbol{Y_{\text{-}i}}, f(\boldsymbol{D})] \mid \boldsymbol{\xi} \tag{24}$$

Under these assumptions, the observed data likelihood can be written as

$$L(\boldsymbol{Y_o}, \boldsymbol{D} \mid f(\boldsymbol{D}); \boldsymbol{\iota}, \boldsymbol{\varphi}) \propto P(\boldsymbol{D} \mid \boldsymbol{Y_o}, f(\boldsymbol{D}); \boldsymbol{\varphi}) \prod_{n=1}^{N} \int_{\mathbb{R}^k} P\left(\boldsymbol{Y_o}^{(n)} \mid \boldsymbol{\xi}; \boldsymbol{\iota}\right) g(\boldsymbol{\xi} \mid f(\boldsymbol{D_n}); \boldsymbol{\iota}) \, d\boldsymbol{\xi}. \tag{25}$$

Hence, missingness does not depend on unobserved variables in the model, and the two factors of the likelihood function can be maximized independently. The final MML estimator of the LRM for nonignorable item nonresponses is then

$$L(\boldsymbol{Y_o} \mid f(\boldsymbol{D}); \boldsymbol{\iota}, \boldsymbol{\varphi}) \propto \prod_{n=1}^{N} \int_{\mathbb{R}^k} \prod_{i=1}^{I} P(Y_{ni} = y_{ni} \mid \boldsymbol{\xi}; \boldsymbol{\iota})^{d_{ni}} \, g(\boldsymbol{\xi} \mid f(\boldsymbol{D_n}); \boldsymbol{\iota}) \, d\boldsymbol{\xi}. \tag{26}$$

The missingness information with respect to persons' latent trait levels are taken into account by the conditional distribution $g(\boldsymbol{\xi} \mid f(\boldsymbol{D}_n); \boldsymbol{\iota})$ in the model. In most implementations of MML estimators, a conditional multivariate normal distribution of the latent trait is assumed, so that

$$\boldsymbol{\zeta} \sim MVN(\boldsymbol{0}, \Sigma_\zeta) \tag{27}$$

where $\boldsymbol{\Sigma}_\zeta$ is the covariance matrix of the residuals $\boldsymbol{\zeta} = \boldsymbol{\xi} - E[\boldsymbol{\xi} \mid f(\boldsymbol{D})]$. The model can easily be extended by inclusion of additional covariates $\boldsymbol{Z}$ in the LRM. The general final model is graphically represented in Figure 2.

For applications of the LRM it is crucial to find appropriate functions $f(\boldsymbol{D})$. This might be easy in some cases. For example, if the test design does not allow for omissions of items, so that nonresponses only result from not-reached items, all information of the missing pattern $\boldsymbol{D}$ is given by the number of reached or not-reached items, which is S_D or $I - S_D$.

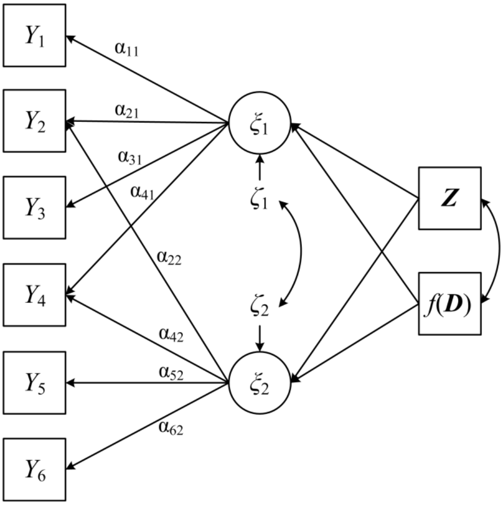

Figure 2:
The measurement model of ξ with the function $f(\boldsymbol{D})$ and additional covariates **Z** as predictors in the LRM for nonignorable item nonresponses

There are many conceptual similarities between MIRT models and LRMs, but there are also significant differences. In MIRT models local stochastic independence is assumed for the response indicators D_i (see Equation 19). This assumption is necessarily violated in the case of not-reached items (Rose, 2013; Rose et al., 2015). In LRMs for nonignorable missing responses the assumption of local stochastic independence with respect to response indicators D_i is not required. For that reason MIRT models are appropriate to account for omitted response, whereas LRMs are recommended to account for item nonresponses due to not-reached items (Rose, 2013; Rose et al., 2015).

In some applications the choice of appropriate functions $f(\boldsymbol{D})$ is difficult. We recommend an analysis of $\boldsymbol{D}$ in a first step. If the assumption of a latent responses propensity underlying $\boldsymbol{D}$ is tenable, estimates of $\boldsymbol{\theta}$, such as ML or EAP estimates, can be used as functions $f(\boldsymbol{D})$ in the latent regression model. Note that the simplest function $f(\boldsymbol{D})$ is $\boldsymbol{D}$ itself. Hence, the latent regression can also be specified as the multiple regression $E(\boldsymbol{\xi} \mid D_1, \ldots, D_I)$; even interaction effects between the response indicators are allowed but may rapidly result in large numbers of independent variables in the LRM.

A potential disadvantage of using LRMs is the possible unreliability in functions $f(\boldsymbol{D})$ especially if estimates of $\boldsymbol{\theta}$ obtained from a previously fitted model of $\boldsymbol{D}$ are used. However, Rose (2013) showed that even in cases with a minimal number of items and very low reliabilities of EAP estimates ($\mathrm{Rel}\left(\hat{\theta}_{EAP}\right)$ = .41 − .55), the LRM equivalently adjusted for item nonresponses compared to more complex MIRT models.

Multiple group models for nonignorable item nonresponses

MG-IRT models for nonignorable item nonresponses are conceptually very close to LRMs. The basic idea is to use categorical functions $f(\boldsymbol{D})$ that serve as grouping variables in a multiple group IRT model. Let $q = 1, \ldots ,Q$ be the values of $f(\boldsymbol{D})$. The MML estimation equation of the MG-IRT model for missing data is

$$L\left(\boldsymbol{Y}_o \mid f\left(\boldsymbol{D}\right);\boldsymbol{\iota},\boldsymbol{\varphi}\right) \propto \prod_{q=1}^{Q}\prod_{n=1}^{N_q}\int_{\mathbb{R}^k}\prod_{i=1}^{I} P\left(Y_{ni} = y_{ni} \mid \boldsymbol{\xi};\boldsymbol{\iota}\right)^{d_{ni}} g\left(\boldsymbol{\xi} \mid f\left(\boldsymbol{D}_n\right) = q;\boldsymbol{\iota}\right) d\boldsymbol{\xi}. \quad (28)$$

The assumptions in the MG-IRT model are equivalent to that of the LRM model discussed in the previous section (see Equation 23 and 24). Since no DIF is assumed with respect to $f(\boldsymbol{D})$, the item parameters are constrained to be equal across the groups. The distribution of the latent variables in the groups is assumed to be multivariate normal, with

$$\boldsymbol{\xi} \mid f\left(\boldsymbol{D}\right) = q \sim MVN\left(E\left[\boldsymbol{\xi} \mid f\left(\boldsymbol{D}\right) = q\right], \boldsymbol{\Sigma}_{\xi|q}\right) \quad (29)$$

In contrast to the LRM, not only are the means allowed to differ across the values of $f(\boldsymbol{D})$ but the covariance matrices $\boldsymbol{\Sigma}_{\xi|q}$ can be group specific as well.

As in the case of LRMs, the choice of the functions $f(\boldsymbol{D})$ is crucial. Theoretically, each missing pattern can be considered a separate group. In this case the MG-IRT model is equivalent to a PMM with additional assumptions (see Equations 23 and 24). Unfortunately, the number of missing data patterns and groups, respectively, are typically very large and alternative functions $f(\boldsymbol{D})$ should be used. For example, Rose et al. (2010) used stratified students of the PISA 2006 sample into three strata depending on their response rates and applied a three group IRT model. They found substantial mean differences in the latent mathematics, science, and reading proficiency levels between the strata, which supports the assumption that omitted and not-reached items in PISA 2006 were most likely nonignorable.

Using information of missingness in MML estimation

Before the different models described here are critically compared, we explain how they adjust for nonignorable nonresponses. We confine ourselves to the case of MML estimation using an EM algorithm as proposed by Bock and Aitkin (1981). In general, the EM algorithm is an iterative procedure consisting of two steps: (a) the E-step where the expectation $E\left[\ell\left(\boldsymbol{Y},\boldsymbol{D} \mid \boldsymbol{Z};\boldsymbol{\iota},\boldsymbol{\varphi}\right)\right]$ of the log-likelihood is calculated, and (b) the M-step, where the expected log-likelihood is maximized. The basic idea of MML estimation is to assume a parametric conditional or unconditional distribution of the latent variables. Typically a multivariate normal distribution is assumed, which is represented by $g(\boldsymbol{\xi}, \boldsymbol{\theta} \mid \boldsymbol{Z}_n; \boldsymbol{\iota})$ in the MIRT model (see Equation 20). This allows for removing individual person parameters from the MML estimation equation by integrating the distribution of the latent variables. This needs to be done for each of the N individual response patterns in the sample.

The individual distributions $g(\boldsymbol{\xi}, \boldsymbol{\theta} \mid \boldsymbol{Z}_n)$ are unknown. However, the Bayes theorem allows for computing the joint posterior distribution of the latent variables given the observed responses, the missing pattern and the covariates. That is

$$g\left(\boldsymbol{\xi},\boldsymbol{\theta} \mid \boldsymbol{Y}_o^{(n)},\boldsymbol{D}_n,\boldsymbol{Z}_n\right)=\frac{g\left(\boldsymbol{\xi},\boldsymbol{\theta},\boldsymbol{Y}_o^{(n)},\boldsymbol{D}_n \mid \boldsymbol{Z}_n\right)}{\int_{\mathbb{R}^k}\int_{\mathbb{R}^p} g\left(\boldsymbol{\xi},\boldsymbol{\theta},\boldsymbol{Y}_o^{(n)},\boldsymbol{D}_n \mid \boldsymbol{Z}_n\right)d\boldsymbol{\xi}d\boldsymbol{\theta}} \tag{30}$$

Under the assumptions of the MIRT model (see Equations 18 and 19) this can be written as

$$g\left(\boldsymbol{\xi},\boldsymbol{\theta} \mid \boldsymbol{Y}_o^{(n)},\boldsymbol{D}_n,\boldsymbol{Z}_n\right)=\frac{P\left(\boldsymbol{Y}_o^{(n)} \mid \boldsymbol{\xi}\right)g\left(\boldsymbol{\xi} \mid \boldsymbol{\theta},\boldsymbol{Z}_n\right)P\left(\boldsymbol{D}_n \mid \boldsymbol{\theta}\right)g\left(\boldsymbol{\theta} \mid \boldsymbol{Z}_n\right)}{\int_{\mathbb{R}^k}\int_{\mathbb{R}^p} g\left(\boldsymbol{\xi},\boldsymbol{\theta},\boldsymbol{Y}_o^{(n)},\boldsymbol{D}_n \mid \boldsymbol{Z}_n\right)d\boldsymbol{\xi}d\boldsymbol{\theta}} \tag{31}$$

Finally, the posterior density of the latent variable $\boldsymbol{\xi}$ of observation n is

$$g\left(\boldsymbol{\xi} \mid \boldsymbol{Y}_o^{(n)},\boldsymbol{D}_n,\boldsymbol{Z}_n\right)=\frac{P\left(\boldsymbol{Y}_o^{(n)} \mid \boldsymbol{\xi}\right)\int_{\mathbb{R}^p} g\left(\boldsymbol{\xi} \mid \boldsymbol{\theta},\boldsymbol{Z}_n\right)P\left(\boldsymbol{D}_n \mid \boldsymbol{\theta}\right)g\left(\boldsymbol{\theta} \mid \boldsymbol{Z}_n\right)}{\int_{\mathbb{R}^k}\int_{\mathbb{R}^p} g\left(\boldsymbol{\xi},\boldsymbol{\theta},\boldsymbol{Y}_o^{(n)},\boldsymbol{D}_n \mid \boldsymbol{Z}_n\right)d\boldsymbol{\xi}d\boldsymbol{\theta}} \tag{32}$$

Hence, the test takers' posterior density of the latent competency variable $\boldsymbol{\xi}$ depends not only on the observed item responses but also on the relationship between $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$. If the dimensions ξ_m and θ_l are positively correlated, persons with more item nonresponses have left-shifted posterior distributions of $\boldsymbol{\xi}$ and lower expected values of ξ_m respectively. Similarly, the individual posterior density of $\boldsymbol{\xi}$ in LRMs is

$$g\left(\boldsymbol{\xi} \mid \boldsymbol{Y}_o^{(n)},f\left(\boldsymbol{D}_n\right),\boldsymbol{Z}_n\right)=\frac{P\left(\boldsymbol{Y}_o^{(n)} \mid \boldsymbol{\xi}\right)g\left(\boldsymbol{\xi} \mid f\left(\boldsymbol{D}_n\right),\boldsymbol{Z}_n\right)}{\int_{\mathbb{R}^k} P\left(\boldsymbol{Y}_o^{(n)} \mid \boldsymbol{\xi}\right)g\left(\boldsymbol{\xi} \mid f\left(\boldsymbol{D}_n\right),\boldsymbol{Z}_n\right)d\boldsymbol{\xi}} \tag{33}$$

If the latent competency $\boldsymbol{\xi}$ is related to missingness the posterior distribution is shifted and the expected a posterior proficiency level changes. For example, if the sum S_D is chosen as the function $f(\boldsymbol{D})$ in the LRM, and a positive regression coefficient of $f(\boldsymbol{D})$ indicates nonignorable missing data, then higher trait levels are more likely in observations with high response rates. In turn, observations with low response rates have left-shifted posterior distributions and, therefore, lower expected trait levels. Based on all individual posterior distributions the adjusted expected rate of correct and incorrect responses calculated for each item. These rates are considered in the first and second derivatives of the log-likelihood with respect to the item parameters. Hence, the information of missingness is accumulated in the E-step and taken into account in the M-step when the expected log-likelihood is maximized.

Comparing model-based approaches for nonignorable item nonresponses

In this section, we compare MIRT models, LRMs and MG-IRT models highlighting their commonalities and differences, which have important implications for their suitability in real applications. A common characteristic of all models discussed here is the inclusion of missingness represented by the response indicators D_1 , . . . , D_I. A major difference between MIRT models on the one hand LRMs and MG-IRT models on the other hand is that $\boldsymbol{D}$ is only modeled in MIRT models using a latent variable model. Therefore, Pohl et al. (2015) denoted this approach as the latent approach. LRMs and MG-IRT models can be regarded as conditional models with functions $f(\boldsymbol{D})$ of the manifest responses indicators as covariates in a background model or as grouping factors. Accordingly, this class of models was denoted as the manifest approach (Pohl et al., 2014). The assumptions differ between the two approaches, which has implications for the indication of each method. Only in the latent approach is local stochastic independence assumed for the response indicator variables D_i (see Equation 19). This assumption is not required in LRMs and MG-IRT models. Hence, the manifest approach is preferred when local stochastic independence of response indicators does not hold. This is the case when the missing responses result from not-reached items; the value of D_i depends on whether item $i - 1$ was not reached or item $i + 1$ was reached in time (Rose, 2013; Rose et al., 2014). The use of MIRT models is appropriate to account for omitted responses. Nevertheless, both approaches, the manifest and the latent, rest upon implicit assumptions about how the dependencies between $\boldsymbol{D}$ and $\boldsymbol{Y}$ are implied. In MIRT models it is assumed that the stochastic relationship between ξ and $\boldsymbol{\theta}$ implies all dependencies between items and responses indicators. In the manifest approach, however, the stochastic dependency of ξ on functions $f(\boldsymbol{D})$ implies dependency of $\boldsymbol{Y}$ on $\boldsymbol{D}$. In other words, all information of missingness ($\boldsymbol{D}$) with respect to unobserved variables $\boldsymbol{Y_m}$ and ξ is represented in the structural model of latent variables in a joint model of ($\boldsymbol{Y}$, $\boldsymbol{D}$) (MIRT) or in a model that allows for distributional differences in ξ and $\boldsymbol{Y}$ depending on functions $f(\boldsymbol{D})$ (LRM and MG-IRT). The latent and the manifest approaches are very close, when the assumptions of the latent variable approach hold and proxies of $\boldsymbol{\theta}$ are used as functions $f(\boldsymbol{D})$.

Considering the issue of model equivalence, MIRT models, LRMs, and MG-IRT models are equivalent in the construction of the latent variable ξ. As noted previously, this is a fundamental requirement, since model-based approaches are not intended to change the target model, rather they should adjust for missing data. Furthermore, the three types of models under comparison are approximately equivalent in adjusting for missing responses when estimates $\hat{\boldsymbol{\theta}}$ are used in an LRM despite the unreliability in person parameters estimates of the latent response propensity. MG-IRT models and LRMs are conceptually identical if $f(\boldsymbol{D})$ is categorical, but differ in the assumptions about the distribution of the residual $\boldsymbol{\zeta}$. The LRM assumes equality of variances and covariances for all values of $f(\boldsymbol{D})$, whereas MG-IRT models allow for heterogeneity in the latent residuals. This difference may guide researchers in real applications to choose between the models.

How to choose the appropriate model

From the theoretical considerations we can derive a set of questions that needs to be answered prior to analyses to choose the right model. In this section we will list these questions and provide answers that may guide researchers in real applications. We begin with the first and most important question about the missing data mechanism and will end with concrete questions about model specification.

Is there really a problem with item nonresponses? It is important to keep in mind that the problem of missing data is negligible if the nonresponse rate is very low, even if the missing data mechanism is strongly NMAR. However, the use of IRT models in LSAs aims at accurate estimation of item parameter and distribution parameters of the latent variables in all subpopulations of interest. Therefore, the rates of missing responses should be examined for all items and in all subpopulations addressed in the study. If only small rates of missing data are found, models for nonignorable missing data are not required. This is all the more true since the common MML estimator used for IRT models is a FIML estimator. Hence, all observed item responses are used and validity of the MAR assumption is sufficient for unbiased parameter estimation.

Is the missing data mechanism NMAR? This question is essential, since an ignorable missing data mechanism implies that $\boldsymbol{D}$ can be left out. Unfortunately, the question cannot be tested or answered empirically. So, the question is how plausible is an ignorable missing data mechanism. The test design may have implications for the plausibility and tenability of the ignorability assumption. There might also be hints in the data suggesting nonignorability, for example, rates of omitted or not-reached items that are stochastically dependent on proportion correct scores of the observed item responses (Rose et al., 2010). Similarly, missing rates per item may depend on the observed item means, as an inverse measure of their difficulty. Such findings indicate that missingness is systematic and therefore MAR or even NMAR. If the ignorability assumption is implausible an appropriate model of $\boldsymbol{Y}$ and $\boldsymbol{D}$ needs to be chosen.

Are nonresponses the result of omitted or not-reached items? In general, MIRT models discussed in this paper are appropriate for omitted responses, whereas nonignorable missing values due to not-reached items are better handled by LRMs (Rose, 2013; Rose et al., 2015) or special MIRT models with additional constraints (Glas & Pimentel, 2008; Pohl et al., 2014). If missing values in $\boldsymbol{Y}$ result from omitted and not-reached items, LRMs and MIRT models can be combined (Rose, 2013; Rose et al., 2015) or MIRT models with separate latent response propensities for omitted and not-reached items can be used (Pohl et al., 2014).

As in any latent variable models it is recommended to establish the model gradually. Thus, if the assumption of local stochastic independence of response indicators is justifiable and a latent responses propensity can be assumed, the structure of $\boldsymbol{D}$ should be examined in a first step to answer the question: *What is the best model for* $\boldsymbol{D}$? Exploratory factor analyses for dichotomous variables can be used in this step (Wirth & Edwards, 2007) as well as other techniques to investigate dimensionality (Reckase, 2009). Even if an LRM is used it is strongly advised to analyze $\boldsymbol{D}$ initially to choose appropriate functions $f(\boldsymbol{D})$ for the LRM or the MG-IRT model. If many dimensions and a complex factorial structure are needed to model $\boldsymbol{D}$, it is best to use a two-step procedure. Estimates of $\hat{\boldsymbol{\theta}}$ are obtained in the first step form a model of $\boldsymbol{D}$, which are used as functions $f(\boldsymbol{D})$ in a LRM for nonignorable item nonresponses in the second step.

As previously noted, MIRT models can only account for linear relationships between latent propensities and latent traits. Therefore, applied researchers need to answer the question: *Is the relationship between dimensions θ_l and ξ_m linear?* To answer this question estimates $\hat{\theta}_l$ and polynomials $\hat{\theta}_l^h$, with $h = 1, \ldots, H$, can be used in an LRM to investigate potential nonlinearity. If there is strong evidence for violations of the linearity assumption, LRMs are preferred.

What is the appropriate function $f(\boldsymbol{D})$ for an LRM or an MG-IRT model? This question needs to be answered if the manifest approach is chosen. As previously mentioned, the analysis of $\boldsymbol{D}$ itself can help to answer this question. For the case of omitted responses IRT person parameter estimates $\hat{\boldsymbol{\theta}}$ might be preferred if the dimensionality underlying $\boldsymbol{D}$ is complex. As the sum score is a sufficient statistic for the latent trait in the Rasch model, the response rates S_D can be used in an LRM if $\boldsymbol{D}$ can be appropriately modeled by a unidimensional Rasch model. This is also possible if $\boldsymbol{\theta}$ is a p-dimensional latent response propensity in a multidimensional Rasch model with simple structure. In this case each D_{il} indicates only one latent dimension θ_l. The p sum scores $S_{Dl} = \sum_{i=1}^{I_l} D_{il}$ can be used in a latent regression $E\left(\xi \mid S_{D1}, \ldots, S_{Dp}\right)$.

If a simple structure does not hold for the measurement model of $\boldsymbol{\theta}$ or if a one-parameter model may be inappropriate, the IRT person parameter estimates may be used rather than sum scores. Even if the Rasch model holds for $\boldsymbol{\theta}$ in assessments with incomplete booklet designs, the sum scores S_D or S_{Dl} may not be comparable across booklets since they result from different response indicator variables depending on the test forms. In such cases IRT person parameter estimates are also superior to observed responses rates. LRMs are very flexible and additionally allow for interactions between test forms and functions $f(\boldsymbol{D})$ with respect to $\boldsymbol{\xi}$, by inclusion of indicator variables of the test forms and respective interactions terms. Therefore, the test design also needs to be considered in deciding which missing data model is appropriate and how should it be specified.

If the number of manifest and latent variables is large and the sample size moderate to small LRMs and MG-IRT models might be preferred. The same is true in IRT models for longitudinal data. In this case the latent responses propensity needs to be modeled for each time point in MIRT models for nonignorable item nonresponses. Model complexity might then become impractical even for large sample sizes and two-step modeling might be a better choice. $\boldsymbol{D}$ is examined in a first step to find an appropriate measurement model of $\boldsymbol{\theta}$ and appropriate functions $f(\mathbf{D})$.

What degree of model complexity is compatible with the sample size and the number of items? Model complexity is an important issue in selecting the most appropriate model in a concrete application. If appropriate functions $f(\boldsymbol{D})$ can be found LRMs and MG-IRT models are often less complex and computationally less demanding. Especially in cases with multidimensional latent traits and multidimensional latent response propensities, MIRT models may become inapplicable. For example in IRT models for longitudinal data the latent responses propensity needs to be modeled for each time point. LRMs and MG-IRT models with $f(\boldsymbol{D}) = \hat{\boldsymbol{\theta}}$ can be a viable alternative. However, this question needs to be answered considering the number of cases and the number of items in the study. More complex models with more parameters require larger samples. Unfortunately, no clear recommendations can be given for required sample sizes. Monitoring convergence and comparing parameter estimates and standard errors obtained from different model-based approaches may facilitate decision-making for an appropriate and stable model.

Are there potential moderating variables that need to be taken into account? In both, the latent and the manifest approach for nonignorable item nonresponses, covariates $\boldsymbol{Z}$ can be included in a background model. However, these covariates can moderate the relationship between $f(\boldsymbol{D})$ and $\boldsymbol{\xi}$ or $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$. In the manifest approach, interaction terms can simply be included in the LRM. Existing IRT software hardly allows for interactions between manifest and latent variables. However, if the covariates are categorical with a few levels, MG-MIRT models can be used with equally constraint parameters of the measurement model of $\boldsymbol{\xi}$ across groups.

How to deal with the fact the missing data mechanism within single items can be different? Missing responses may result not only from omitted or not-reached items but also from the test design. Planned missing data due to not administered items are typically MCAR if different test forms with only subsets of items were randomly assigned to test takers. Omitted and not-reached items, however, may be MAR or even NMAR. This implies that different missing data mechanisms can coexist. Since, all models described in this paper adjust for nonignorable missing data, missingness that is MAR or MCAR should not be addressed by these approaches, respectively. This can be achieved by defining response indicators more specifically. Instead of indicating an observed response ($D_i = 1$) or an item nonresponse ($D_i = 0$), they can be defined as response indicators of administered items only. In this case, we do not know whether an item *i* would have been answered by test takers or not if *i* were administered. Hence, $\boldsymbol{D}$ itself is incomplete in cases of planned missing data. This is not problematic as long as the missing data mechanism is MCAR. Nevertheless, the analysis of $\boldsymbol{D}$ and finding appropriate functions $f(\boldsymbol{D})$ can become more difficult with incomplete data matrices $\boldsymbol{D} = \boldsymbol{d}$. Similarly, item nonresponses due to omitted and not-reached items can differ with respect to the missing data mechanism. In this case two sets of indicator variables can be defined. The first set indicates omitted ($D_i^{(R)} = 0$) or not omitted items ($D_i^{(R)} = 1$) and the second set indicates reached ($D_i^{(R)} = 1$) or not-reached items ($D_i^{(R)} = 0$). Both can be modeled separately (Pohl et al., 2014; Rose, 2013; Rose et al., 2015).

Note that this list of questions is by no means complete but can make researchers aware of the most relevant issues in model selection and may provide at least some guidance in IRT scaling with potentially nonignorable item nonresponses.

Discussion

Missing responses may occur for many different reasons. In educational and psychological assessments item nonresponses are often not under researchers' control. Missing responses due to omitted and not-reached items typically occur systematically, implying that the missing data mechanism is MAR or even NMAR. In the missing data literature many different approaches have been developed that account for ignorable item nonresponses but just a few approaches exists for nonignorable missing data. In this paper, we initially adopted the definitions of Rubin's missing data mechanisms to the peculiarities of educational and psychological measurement. By means of these definitions we considered ML estimation with missing data in IRT models. A general MML estimator was derived, which underlies existing MIRT models for nonignorable missing data. We further compared MIRT models, LRMs, and MG-IRT models regarding their commonalities and differences and discussed the implications for real applications. MIRT models assume a latent response propensity variable underlying $\boldsymbol{D}$ and are denoted as the latent approach (Pohl et al., 2014). LRMs and the MG-IRT models refer to the manifest approach, since they do not need the assumption of a latent variable underlying $\boldsymbol{D}$ (Pohl et al., 2014). Instead, information of missingness is taken into account using functions $f(\boldsymbol{D})$ as predictors in an LRM or as grouping factors in MG-IRT models.

Comparisons of models should be based on certain criteria. Model equivalence is a widely used concept to judge equality of nested models. Considering that missing data models aims to adjust for nonresponses, we extended the concept of model equivalence. In most cases it is sufficient to define model equivalence as the equality of model-implied distributions of manifest variables and equal goodness-of-fit statistics. However, in missing data models the equivalent construction of the latent trait variable based on $\boldsymbol{Y}$ and the equality in adjusting for missing values are two additional and even more important aspects of model equivalence. In the latent approach, different MIRT models exist, which are equal in terms of all three aspects of model equivalence. LRMs and MG-IRT include different sets of variables and parameters than MIRT model. Hence, they are not nested and incomparable in terms of model fit, respectively. Nevertheless, all models are equivalent in the construction of the latent trait ξ, which means that the measurement model of $\boldsymbol{Y}$ remains an unchanged part in all models. LRMs and MIRT models adjust equally well for item nonresponses if person parameter estimates of the latent responses propensities are used as predictors in the LRM (Rose, 2013). Furthermore, the MG-IRT model and the LRM with dummy variables are equivalent if the predictor $f(\boldsymbol{D})$ is categorical and homogeneity with respect to ξ holds for all levels of $f(\boldsymbol{D})$.

The models of the manifest approach were originally developed as less complex alternatives for MIRT models, implying that they have a lot in common. This is true if the proxies of a latent response propensity serve as predictors in an LRM. However, the latent and the manifest approach differ in their assumptions. Local stochastic independence of response indicators is assumed in MIRT models only. This assumption can either strongly be violated or the assumption of a latent response propensity measured by $\boldsymbol{D}$ may not be justifiable. In such situations LRMs or MG-IRT models are the method of choice (Rose, 2013; Rose et al., 2015). LRMs and MG-IRT models assumes that all information of $\boldsymbol{D}$ with respect to unobserved variables in the model is preserved in functions $f(\boldsymbol{D})$. Hence, finding appropriate functions $f(\boldsymbol{D})$ which met this assumption is crucial. However, in some cases these functions are easy to find, for example if item responses result from not-reached items in a timed test and all test takers answered the same items in the same order. All information of missingness is given by the number of reached (S_D) or not-reached items ($I - S_D$). In most cases the choice of functions $f(\boldsymbol{D})$ is not straightforward. We strongly recommend to analyze $\boldsymbol{D}$ prior to the final scaling either to find the correct dimensionality of the latent response propensity or to find suited functions $f(\boldsymbol{D})$ for the manifest approach. LRMs and MG-IRT models have the disadvantage of finding appropriate functions $f(\boldsymbol{D})$ but have the advantage of being very flexible. Nonlinearity between ξ and $f(\boldsymbol{D})$ can easily be taken into account, as well as interactions between additional covariates and $f(\boldsymbol{D})$. Hence, the models of the manifest approach are promising for a wide range of applications.

The aim of this paper is not only to compare the different models that have been developed, but to provide some insights about how these models correct for missing responses. Unfortunately, this is specific for different estimators and we constrained our considerations to MML estimation using the EM algorithm. However, we could show that the adjustment mechanism is very similar in the latent and the manifest approach. This paper is also intended to provide some guidance in selection and application of models for nonignorable item nonresponses. For that reason we used a list of questions that should be addressed by applied researchers in order to select a particular model in a concrete application. Unfortunately, some of the recommendations have remained vague for several reasons. First, although MIRT models for missing responses were developed more than a decade ago they are still not widely applied. Hence, there is a lack of practical knowledge in the use of models for nonignorable missing data. Second, some of the questions may never be definitely answered in many applications, such as the question about the missing data mechanism of unplanned item nonresponses. Testing the nonresponse mechanism would require observing the missing part $\boldsymbol{Y}_m$ and the latent trait in the model. Hence, knowledge must be replaced by plausibility considerations about certain assumptions. For the sake of simplicity, it is commonly assumed that the missing data mechanism is MAR because then missingness does not need to be included in the model. However, item nonresponses should be taken seriously even if missingness is ignorable. We showed that three different MAR assumptions can be differentiated in latent trait models. Two of them require the appropriate inclusion of covariates in parameter estimation. This is well known for the case of SEM (Graham, 2003) but hardly discussed in IRT modeling (e. g. DeMars, 2002).

The missing data mechanism may be termed ignorable or nonignorable, but that does not mean that applied researchers can ever ignore missing values. Rather they need to decide how to deal with them. Here we summarized and compared different models which reduce nonresponse biases because of nonignorable missing responses. However, they rest on fairly strong assumptions. Violations of these assumptions can result in more biased parameter estimates than using models which ignore missingness. Sensitivity analyses have been proposed as a means for investigating dependence of parameter estimation on model assumptions. In the discussed IRT models for nonignorable item nonresponses, especially the conditional stochastic independence assumptions (see Equations 18, 19, 23, and 24) are critical. Unfortunately, violations of these assumptions may not only affect parameter estimation but pose theoretical challenges. For example, if items Y_i are not conditionally stochastically independent of the response indicators $\boldsymbol{D}$ or functions $f(\boldsymbol{D})$, differential item functioning due to the response indicators (or functions of it) is implied. The theoretical and practical implications of such violations are still unclear and should be addressed in future research. Furthermore, guidance for sensitivity analyses should be developed, which also accounts for the amount and the complexity of data in LSA. Finally, the development of less restrictive models with fewer assumptions is an objective of future research.

Author note

Parts of this paper are based on the unpublished dissertation of the first author.

We thank Andreas Frey and Rolf Steyer who served as members on the thesis committee.

We are very grateful to Robert J. Mislevy and Shelby J. Haberman from the Educational Testing Service (Princeton, US), for their insightful comments and advice.

References

Adams, R., Dumais, J., Foy, P., Hastedt, D., Kelly, D. L., Macaskill, G., . . . Shen, C. (1998). *International Association for the Evaluation of Educational Achievement User Guide for the TIMSS International Database* (E. J. Gonzalez, T. A. Smith, & H. Sibberns, Eds.). Retrieved from http://timss.bc.edu/timss1995i/database/UG3.pdf.

Asparouhov, T., & Muthén, B. O. (2012). *Comparison of computational methods for high dimensional item factor analysis* (Mplus Webnote). Los Angeles, CA: Muthén & Muthén.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). Boca Raton, FL: CRC. doi: 10.2307/2532822 Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46, 443–459. doi: 10.1007/BF02293801

Cai, L. (2010). Metropolis-hastings robbins-monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, 35(3), 307–335. doi: 10.3102/1076998609353115

Culbertson, M. (2011, April). *Is it wrong? Handling missing responses in IRT*. Speech presented at the Annual Meeting of the National Council on Measurement in Education. New Orleans, LA.

DeMars, C. (2002). Incomplete data and item parameter estimates under JMLE and MML. *Applied Measurement in Education*, 15, 15–31. doi: 10.1207/S15324818AME1501_02

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement*, 13(1), 77–90. doi: 10.1177/014662168901300108

Glas, C. A. W. (2006). *Violations of ignorability in computerized adaptive testing* (Research Report No. Report 04-04). Enschede, The Netherlands: University of Twente.

Glas, C. A. W., & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement*, 68(6), 907–922. doi: 10.1177/0013164408315262

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 10(1), 80–100. doi: 10.1207/S15328007SEM1001_4

Graham, J. W., Taylor, B., Olchowski, A., & Cumsille, P. (2006). Planned missing data designs in psychological research. *Psychological Methods*, 11(4), 323–343. doi: 10.1037/1082-989X.11.4.323

Graham, J. W., Taylor, B. J., & Cumsille, P. E. (2001). Planned missing-data designs in analysis of change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 335–353). Washington, DC, US: American Psychological Association.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *The Annals of Economic and Social Measurement*, 5, 475-492.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153–61. doi: 10.2307/1912352

Holman, R., & Glas, C. A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, 58(1), 1–17. doi: 10.1111/j.2044-8317.2005.tb00312.x

Korobko, O. K., Glas, C. A. W., Bosker, R. J., & Luyten, J. W. (2008). Comparing the difficulty of examination subjects with item response theory. *Journal of Educational Measurement*, 45, 137–155. doi: 10.1111/j.1745-3984.2007.00057.x

Li, L., Shen, C., Li, X., & Robins, J. M. (2011). On weighting approaches for missing data. *Statistical Methods in Medical Research*, 22(1), 14–30. doi: 10.1177/0962280211403597

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88, 125–134. doi: 10.2307/2290705

Little, R. J. A. (2008). Selection and pattern-mixture models. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), Longitudinal data analysis (p. 409-432). Chapman & Hall/CRC Press.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley & Sons.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, 23(2), 157–162. doi: 10.1111/j.1745-3984.1986.tb00241.x

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement*, 11(1), 81–91. doi: 10.1177/014662168701100106

Mislevy, R. J. (1988). Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters. *Applied Psychological Measurement*, 12(3), 281–296. doi: 10.1177/014662168801200306

Mislevy, R. J., & Wu, P. K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (Research Report No. RR-96-30). 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. doi: 10.1111/1467-985X.00177

OECD. (2009). *PISA 2009 technical report*. Paris: OECD Publishing.

O'Muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: A latent variable approach to item non-response in attitudes scales. *Journal of the Royal Statistic Society*, 162, 177–194. doi: 10.1111/1467-985X.00129

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement*, 74, 423–452. doi: 10.1177/0013164413504926

Puhani, P. (2000). The Heckman correction for sample selection and its critique. *Journal of Economic Surveys*, 14(1), 53–68. doi: 10.1111/1467-6419.00104

Raykov, T., & Marcoulides, G. A. (2001). Can there be infinitely many models equivalent to a given covariance structure model? *Structural Equation Modeling: A Multidisciplinary Journal*, 8(1), 142–149. doi: 10.1207/S15328007SEM0801_8

Raykov, T., & Penev, S. (1999). On structural equation model equivalence. *Multivariate Behavioral Research*, 34(2), 199–244. doi: 10.1207/S15327906Mb340204

Reckase, M. D. (2009). *Multidimensional item response theory*. London, England: Springer. doi:10.1007/978-0-387-89976-3.

Rose, N. (2013). *Item nonresponses in educational and psychological measurement*. (Doctoral thesis, Friedrich-Schiller-University, Jena, Germany). Retrieved from: http://d-nb.info/1036873145/34

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with IRT* (Research Report No. RR-10-11). Princeton, NJ: Educational Testing Service.

Rose, N., Von Davier, M., & Nagengast, B. (2015). *Modeling omitted and not-reached items in IRT models*. Manuscript submitted for publication.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592. doi: 10.1093/biomet/63.3.581

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London, England: Chapman & Hall. doi: 10.1201/9781439821862

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147–177. doi: 10.1037/1082-989X.7.2.147

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, 70(3), 533–555. doi: 10.1007/s11336-003-1141-x

Seaman, S., Galati, J., Jackson, D., Carlin, J., et al. (2013). What is meant by "missing at random"? *Statistical Science*, 28(2), 257–268.

Stelzl, I. (1986). Changing a causal hypothesis without changing the fit: Some rules for generating equivalent path models. *Multivariate Behavioral Research*, 21, 309–331. doi: 10.1207/s15327906mbr2103_3

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, 12(1), 58–79. doi: 10.1037/1082-989X.12.1.58

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, 141(2), 1281–1301. doi: 10.1016/j.jeconom.2007.02.002

Appendix

Derivation MML estimation with item nonresponses. In section three, the observed data MML estimator in presence of item nonresponses was presented (see Equation 14). Here we derived this equation starting with the complete data likelihood (see Equation 13). Using the partition of ***Y*** into the observed ($\boldsymbol{Y_o}$) and the missing part ($\boldsymbol{Y_m}$), we obtain the complete data likelihood

$$L\left(\boldsymbol{Y}_{\boldsymbol{o}},\boldsymbol{D}\mid\boldsymbol{Z};\boldsymbol{\iota},\boldsymbol{\varphi}\right)\propto\prod_{n=1}^{N}\int_{\mathbb{R}^k}\int_{\Omega_m}P\left(\boldsymbol{D}_n\mid\boldsymbol{Y}_{\boldsymbol{o}}^{(n)},\boldsymbol{Y}_{\boldsymbol{m}}^{(n)},\xi,\boldsymbol{Z}_n;\boldsymbol{\varphi}\right)P\left(\boldsymbol{Y}_{\boldsymbol{o}}^{(n)},\boldsymbol{Y}_{\boldsymbol{m}}^{(n)}\mid\xi;\boldsymbol{\iota}\right)g\left(\xi\mid\boldsymbol{Z}_n;\boldsymbol{\iota}\right)d\Omega_m d\xi \tag{34}$$

The assumption of local stochastic independence (see Equation 12) allows us to separate the likelihood of the missing part from the likelihood of the observed part.

$$L\left(\boldsymbol{Y}_{\boldsymbol{o}},\boldsymbol{D}\mid\boldsymbol{Z};\boldsymbol{\iota},\boldsymbol{\varphi}\right)\propto\prod_{n=1}^{N}\int_{\mathbb{R}^k}\int_{\Omega_m}P\left(\boldsymbol{D}_n\mid\boldsymbol{Y}_{\boldsymbol{o}}^{(n)},\boldsymbol{Y}_{\boldsymbol{m}}^{(n)},\xi,\boldsymbol{Z}_n;\boldsymbol{\varphi}\right)P\left(\boldsymbol{Y}_{\boldsymbol{o}}^{(n)}\mid\xi;\boldsymbol{\iota}\right)P\left(\boldsymbol{Y}_{\boldsymbol{m}}^{(n)}\mid\xi;\boldsymbol{\iota}\right)$$
$$g\left(\xi\mid\boldsymbol{Z}_n;\boldsymbol{\iota}\right)d\Omega_m d\xi \tag{35}$$

Due to conditional independence $\boldsymbol{Y_o}\perp\boldsymbol{Y_m}\mid\boldsymbol{\xi}$ the observed part can be brought out of the integral over $\boldsymbol{Y_m}$ (Mislevy & Wu, 1996) yielding

$$L\left(\boldsymbol{Y}_o, \boldsymbol{D} \mid \boldsymbol{Z}; \boldsymbol{\iota}, \boldsymbol{\varphi}\right) \propto \prod_{n=1}^{N} \int_{\mathbb{R}^k} P\left(\boldsymbol{Y}_{\boldsymbol{o}}^{(n)} \mid \xi; \boldsymbol{\iota}\right) \int_{\Omega_m} P\left(\boldsymbol{D}_n \mid \boldsymbol{Y}_{\boldsymbol{o}}^{(n)}, \boldsymbol{Y}_{\boldsymbol{m}}^{(n)}, \xi, \boldsymbol{Z}_n; \boldsymbol{\varphi}\right) \\ P\left(\boldsymbol{Y}_{\boldsymbol{m}}^{(n)} \mid \xi; \boldsymbol{\iota}\right) g\left(\xi \mid \boldsymbol{Z}_n; \boldsymbol{\iota}\right) d\Omega_m d\xi. \tag{36}$$

which is the observed data likelihood given by Equation 14.

In all ignorable missing data mechanisms ***D*** is conditionally stochastically independent off all unobserved variables (***Y******m***, *ξ*) in the model. This allows us to omit the missing part ***Y******m*** from the MML estimation equation. This will be shown here, for the case of the least restrictive ignorable missing data mechanism, which is MAR given (***Y******o***, ***Z***), the MML Equation 36 can be further simplified to

$$L\left(\boldsymbol{Y}_o, \boldsymbol{D} \mid \boldsymbol{Z}; \boldsymbol{\iota}, \boldsymbol{\varphi}\right) \propto \prod_{n=1}^{N} P\left(\boldsymbol{D}_n \mid \boldsymbol{Y}_{\boldsymbol{o}}^{(n)} \boldsymbol{Z}_n; \boldsymbol{\varphi}\right) \int_{\mathbb{R}^k} P\left(\boldsymbol{Y}_{\boldsymbol{o}}^{(n)} \mid \xi; \boldsymbol{\iota}\right) \int_{\Omega_m} P\left(\boldsymbol{Y}_{\boldsymbol{m}}^{(n)} \mid \xi; \boldsymbol{\iota}\right) g\left(\xi \mid \boldsymbol{Z}_n; \boldsymbol{\iota}\right) d\Omega_m d\xi. \tag{37}$$

In Equations 15, 16, and 17 the missing part ***Y******m*** is no longer part of the observed data likelihood since $\int_{\Omega_m} P\left(\boldsymbol{Y}_{\boldsymbol{m}}^{(n)} \mid \xi; \boldsymbol{\iota}\right) = 1$ for all values of the latent trait, and therefore

$$\int_{\mathbb{R}^k} \int_{\Omega_m} P\left(\boldsymbol{Y}_{\boldsymbol{m}}^{(n)} \mid \xi; \boldsymbol{\iota}\right) g\left(\xi \mid \boldsymbol{Z}_n; \boldsymbol{\iota}\right) d\xi d\Omega_m = 1 \tag{38}$$

This result holds also for missing data that MCAR, MAR given ***Y******o***, and MAR given ***Z***.